Global GPU Network (GGN): Performance Benchmarking and Comparative Analysis Against Traditional GPU Computing Models

Rehan Shah[1]*
[1]Ahmedabad International School
*Corresponding Author : rehan.shah@aischool.net

Keywords: Federated Learning, Decentralized Architecture, Blockchain Technology

Abstract

This research presents and evaluates the Global GPU Network (GGN), a distributed computing framework that aims to address GPU accessibility limitations and inequitable digital content monetization by utilizing mobile GPU resources through a decentralized architecture. The study conducted experiments over three consecutive days at three different time intervals (08:00, 14:00, and 00:00) using 10 distributed nodes across Ahmedabad's metropolitan area. The system's performance was compared against a Tesla T4 GPU cloud service using a Convolutional Neural Network (CNN) for Cardiomegaly detection, measuring training duration, operational costs, and network efficiency. The GGN demonstrated average training times of 19 minutes 43 seconds compared to 10 minutes for cloud GPU services, indicating a 97.3% longer processing time. However, the system achieved cost savings of 25-50% during off-peak hours while doubling developer revenue. Performance varied significantly across different time slots, with optimal efficiency during morning (08:00) and night (00:00) sessions. While the GGN shows promise as a cost-effective alternative to traditional GPU solutions, particularly for cost-sensitive applications, its performance variability suggests the need for further optimization. The findings indicate potential for democratizing GPU access while creating new revenue streams for mobile device owners, though technical refinements are needed to improve processing time consistency.

1 Introduction

The rapid expansion of artificial intelligence (AI) applications and cryptocurrency mining operations has resulted in an unprecedented increase in Graphics Processing Unit (GPU) demand, with market analyses indicating price increases of up to 300% (Strzala, 2021). This demand increase, combined with supply chain limitations, has created significant barriers to GPU accessibility, particularly affecting research institutions and small-scale developers. Market projections indicate continued price escalation due to GPUs' essential role in tensor operations and parallel processing capabilities required for emerging technologies (Chen et al., 2023). This trend risks concentrating technological advancement within well-resourced entities, potentially impeding innovation across critical domains including data analytics, AI development, and visual computing.

Concurrent with this GPU accessibility limitation, mobile computing devices constitute an extensive, underutilized computational resource. Current smartphones possess substantial processing capabilities, yet usage patterns demonstrate considerable idle capacity. Research indicates that mobile devices primarily execute low-computational tasks, with social media and messaging applications dominating usage patterns (Li et al., 2022). This underutilization stems from current chip design approaches that optimize for less demanding applications, resulting in significant unused computational potential. The

difference between available processing power and actual utilization presents an opportunity for resource optimization.

The digital content monetization structure presents additional challenges regarding equitable value distribution between platforms and content creators. The prevalent revenue model, exemplified by Google AdSense's 68/32 revenue sharing structure, faces criticism regarding its value distribution mechanics. Although publishers receive 68% of advertising revenue, the platform's 32% retention rate requires examination, particularly considering publishers' operational costs and their role in traffic generation. The transition from cost-per-click (CPC) to cost-per-thousand-impressions (CPM) models has introduced additional complexities, potentially disadvantaging publishers who previously achieved high click-through rates (Varatharajan, 2024).

In response to these interconnected challenges, this research presents the Global GPU Network (GGN), a distributed computing framework that utilizes mobile GPU resources through a decentralized architecture. The GGN implements a combined approach integrating federated learning algorithms with layer-2 blockchain technology to establish a secure, efficient, and economically viable GPU processing network. The framework incorporates an API-based integration system enabling application developers to access distributed GPU resources while generating additional revenue through computational resource sharing.

The system architecture eliminates centralized server infrastructure requirements through the implementation of federated learning protocols and layer-2 blockchain technologies. This approach ensures secure resource allocation and provides transparent compensation mechanisms for participating nodes. The integration of layer-2 blockchain solutions addresses conventional blockchain scalability limitations while maintaining security and decentralization advantages (Shah, 2024).

The research contributions comprise three primary elements: (1) A comprehensive architecture for mobile GPU resource utilization, demonstrating potential for expanded GPU access across technological domains, (2) Implementation and evaluation of a compensation mechanism utilizing layer-2 blockchain technology, providing an alternative revenue model for application developers, and (3) Empirical evidence of system efficiency through network simulations and performance comparisons with traditional GPU solutions.

The research validates GGN effectiveness through empirical evaluation, comparing performance metrics against physical GPUs and cloud-based solutions. The methodology encompasses network simulations under varying load conditions, latency measurements, and cost-benefit analyses. Despite simulation constraints, the experimental design provides insights into operational performance characteristics. These evaluations focus on three metrics: cost efficiency, revenue generation potential, and processing latency.

The distributed node architecture implements load balancing and resource allocation mechanisms. Client nodes facilitate task distribution and coordinate parallel processing activities, while server nodes, comprising mobile devices, execute computational tasks. The implementation includes security protocols and smart contract mechanisms constructed on established layer-2 networks such as Polygon on the Ethereum blockchain, addressing scalability considerations while maintaining security measures

2 Technical Description About Global GPU Network

A comprehensive understanding of the experimental methodology necessitates an examination of the Global GPU Network (GGN) technical architecture and operational framework, as documented in (Shah, 2024) foundational research.

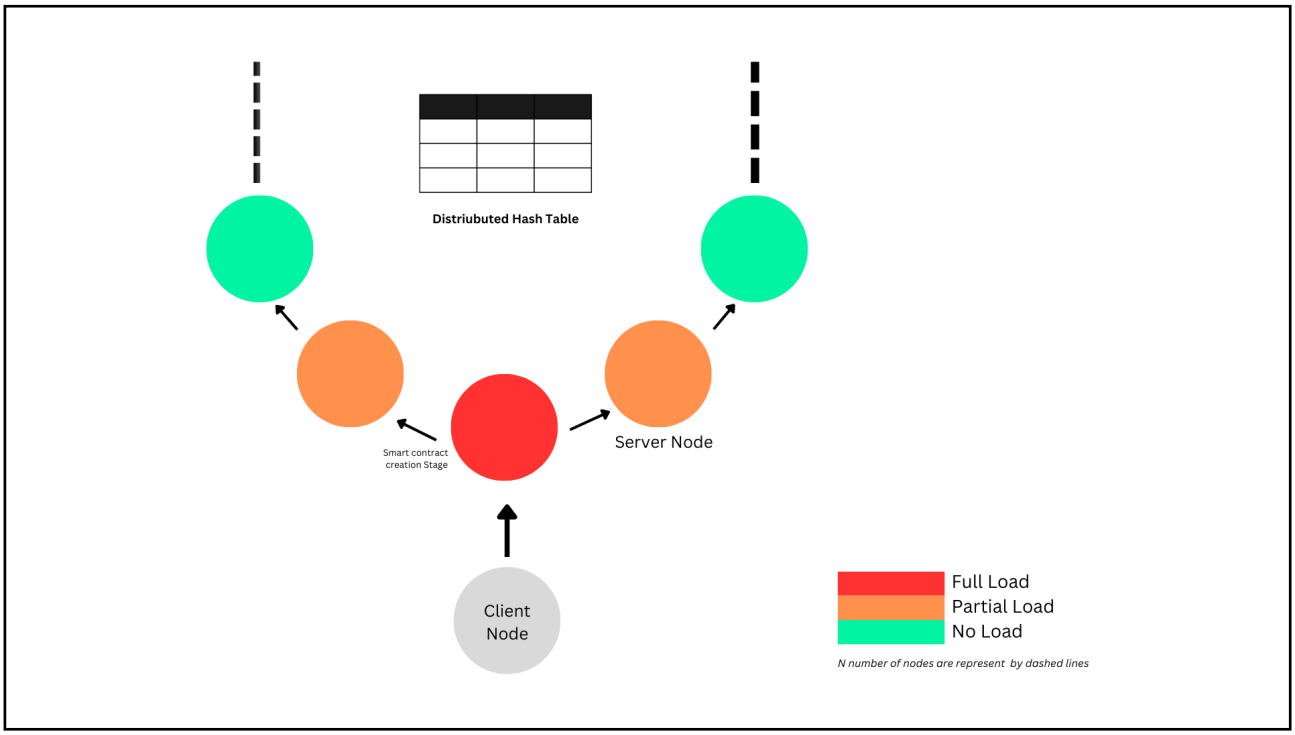

Fig 1: Technical Architecture and Operational Framework of the Global GPU Network (GGN)

The GGN architecture consists of a dual-node system: server nodes and client nodes. The client node operates as the primary distributor, transmitting target information through shared data packets across the network infrastructure (Maymounkov & Mazières, 2002). When capacity thresholds are reached, the system implements the Distributed Hash Table (DHT) to identify proximate nodes with available capacity, executing this process iteratively until resource availability is confirmed.

The network employs a dynamic pricing mechanism to regulate server node participation, adjusting compensation rates based on network demand and computational availability. During high-demand periods, the system increases compensation rates to attract additional nodes, whilst reducing rates during periods of optimal capacity. This equilibrium is governed by:

$$C_t = C_b*(1 + \alpha(\frac{D_t}{S_t} - \beta))$$

Where C_t is the current compensation rate, C_b is the base rate, D_t is network demand, S_t is available capacity, α is price sensitivity, and β represents optimal demand-to-supply ratio. A minimum compensation threshold ensures sustainable network participation during low-demand periods.

The decentralized architecture enables robust performance with minimal maintenance requirements. Network expansion occurs through autonomous propagation: new server nodes transmit their presence to adjacent nodes, enabling automatic table updates and information exchange, thus eliminating centralized server dependencies.

Security protocols incorporate comprehensive end-to-end encryption, including encrypted data storage within individual server nodes. The system addresses potential security vulnerabilities from both server and client nodes, specifically regarding malware introduction. Mitigation measures include Software Development Kit (SDK) implementation for clients, ensuring secure data packet transmission and network offloading. These SDKs require substantial computational resources—though notably less than model training requirements—and enhance security through process sharding and rigorous verification protocols.

As illustrated in Figure 1, the framework demonstrates GGN's capacity for scalable operations with optimized resource utilization. Through distributed, decentralized technology implementation, the network ensures comprehensive accessibility and inclusivity in GPU resource allocation. This architectural framework enables seamless device integration while maintaining efficient and secure computational resource sharing, advancing high-performance computing democratization in artificial intelligence and blockchain environments.

3 Benefits of the Global GPU Network

The research hypothesis postulates that the Global GPU Network (GGN) represents a more economically advantageous and efficient solution for application developers compared to conventional cloud GPU services. This hypothesis emerges from a critical analysis of current market dynamics, characterized by significant monopolistic concentrations in both the GPU manufacturing and cloud computing sectors. Recent antitrust litigation against Google concerning its advertising practices (Rogers, 2024), combined with Nvidia's commanding 90% market share in the GPU sector (Moorhead, 2024), exemplifies the concerning level of market concentration. These monopolistic conditions manifest in exceptional profit margins, as evidenced by Google's extraction of up to 49% of publishers' revenue in numerous cases, and Nvidia's H100 GPU line commanding an unprecedented 1000% profit margin. Such market inefficiencies suggest that the GGN could serve as a disruptive force in democratizing GPU resource allocation.

The research framework acknowledges potential performance trade-offs inherent in distributed systems. Initial analyses indicate that GGN processing times may exceed traditional cloud GPU processing by approximately 20%, attributable to several network-related factors: variability in Wi-Fi bandwidth, network latency fluctuations, and computational overhead associated with smart contract execution. However, this performance differential is expected to diminish as the network achieves scale economies and benefits from increased adoption rates. Furthermore, the distributed nature of the network suggests potential improvements in resource utilization and system redundancy that could partially offset these initial performance constraints.

$H_0: C_G < C_C$

C_G : Revenue from using cloud gpu

C_c : Revenue of using GNN

This primary hypothesis examines the economic efficiency of the GGN compared to traditional cloud services, taking into account both direct costs and indirect benefits of distributed processing.

$H_1: T_G = T_C \times (1 + 0.2)$

T_G : Time used to train using cloud GPU

T_C : Time used to train using GGN

This secondary hypothesis quantifies the expected performance differential between the two systems, incorporating the anticipated 20% processing time increase in GGN operations. This relationship accounts for network-related overhead while maintaining computational integrity.

To empirically validate these hypotheses and assess the viability of the GGN system, a comprehensive simulation environment was constructed. This experimental framework incorporates multiple variables including network latency patterns, bandwidth fluctuations, smart contract execution times, and resource utilization metrics. The simulation aims to provide empirical evidence regarding both the economic advantages and performance characteristics of the proposed system under various operating conditions.

4 Methodology

This experimental investigation evaluates the economic viability and computational efficiency of a Global GPU Network (GGN) as an alternative to traditional cloud GPU services. The study employs temporal variation as the primary independent variable, measuring across three distinct daily intervals (08:00, 14:00, and 00:00) over a three-day period. The dependent variables encompass model training duration and associated operational costs within the GGN environment.

The control experiment implements a Convolutional Neural Network (CNN) for Cardiomegaly detection, utilizing a dataset comprising 1,114 X-ray images trained over 50 iterative cycles to achieve 80% accuracy. The model architecture incorporates 7,515,874 parameters. The code was take from: Rahimanshu. (2023, March). Cardiomegaly Disease Prediction Using CNN, Version 1. Retrieved November 3, 2024, from https://www.kaggle.com/datasets/rahimanshu/cardiomegaly-disease-prediction-using-cnn. The control training environment utilizes a Tesla T4 GPU through immers.cloud, with an hourly operational cost of $0.39. The training duration of 10 minutes resulted in an effective cost of $0.0065. Revenue estimations were derived from average monthly website traffic of 2,000 users (Lindner, 2024), yielding an AdSense revenue of $0.016 per 10-minute interval.

The experimental GGN infrastructure comprises 10 nodes strategically distributed across Ahmedabad's 464.17 km² metropolitan area. The nodes maintain an average separation distance of 3.6 kilometers, a spacing determined through comprehensive network simulation tests optimizing signal strength and interference patterns. The implementation code, available at github.com/Rehan-shah/ggn-sample-code, manages data packet distribution, with each packet containing approximately 75MB of data, encompassing model parameters, training protocols, and validation metrics.

The network infrastructure utilizes iPhone 12 devices, selected based on device lifecycle analysis (Ewa et al., 2017). Each device's Neural Processing Unit (NPU) delivers 11 TOPS of computational capacity (Apple, 2020). To ensure sustainable performance and mitigate thermal throttling, the nodes operate at 30% of maximum capacity, maintaining consistent performance across extended training periods while preventing device degradation.

Given the constrained network scale and single-buyer scenario, the implementation adopts a fixed pricing mechanism rather than dynamic pricing, following Robinson's (1969) market structure analysis. This approach facilitates direct cost calculation and thorough cost-benefit analysis, focusing on creating mutual advantages for developers and GPU providers. The economic framework examines various revenue-sharing scenarios, targeting an equilibrium point where GPU users experience a 50% cost reduction compared to cloud services while generating a 2x increase in developer revenue.

Environmental controls were rigorously maintained throughout the experimental period. Network conditions were standardized with WiFi specifications of 100 ±10 Mbps download and 70 ±6 Mbps upload speeds. Network stability verification preceded each training session, with traffic shaping tools implementing specific bandwidth limitations for each node to ensure uniform speed distribution and prevent network congestion. Docker containerization was employed to standardize the computing environment across varying hardware configurations, ensuring computational consistency despite underlying hardware heterogeneity.

The blockchain implementation utilizes the BNB Smart Chain for transaction processing,

incorporating smart contracts for automated payment distribution and performance verification. The system acknowledges and records transaction cost variations due to inherent blockchain price volatility, maintaining comprehensive documentation of smart contract execution costs across different temporal periods without outlier removal.

The experimental protocol focuses on two primary metrics: the total model training duration, including data distribution and aggregation periods, and the comprehensive costs associated with smart contract creation and execution across different time periods. The experiment executed nine complete training sessions across three daily periods over three consecutive days, capturing temporal variations in both performance and operational costs.

Each experimental session documented complete training cycles, encompassing network initialization, data distribution, model training, and result aggregation phases. This methodological framework enables systematic evaluation of the GGN's operational characteristics while maintaining experimental rigor and reproducibility. The multi-day, multi-period design provides robust comparative data against traditional cloud GPU services, while acknowledging limitations in network scale and geographic distribution that may impact generalizability to larger implementations.

The comprehensive data collection approach ensures capture of all operational aspects, from initial network setup through final result aggregation, providing detailed insights into the practical viability of GGN implementation as an alternative to traditional cloud GPU services. This rigorous methodology enables thorough analysis of both technical performance metrics and economic feasibility factors, essential for evaluating the potential for wider adoption of distributed GPU networks in real-world application

5 Results

The Global GPU Network (GGN) was evaluated through a series of experiments conducted over three consecutive days at three different times each day (09:00, 14:00, and 00:00 IST). The performance was benchmarked against traditional cloud GPU services, specifically using a Tesla T4 GPU. The key performance metrics included model training duration, associated costs, and network latency.

5.1 Model Training Duration

Table 1. Training Duration (in minutes) for GGN and Cloud GPU

Date	Time	GGN Training Duration	Control Training Duration
19th Oct 2024	09:00	18m 42s	10m
19th Oct 2024	14:00	31m 31s*	10m
19th Oct 2024	00:00	17m 13s	10m
20th Oct 2024	09:00	18m 7s	10m
20th Oct 2024	14:00	19m 39s	10m
20th Oct 2024	00:00	16m 25s	10m
21st Oct 2024	09:00	19m 5s	10m
21st Oct 2024	14:00	19m 29s	10m
21st Oct 2024	00:00	17m 14s	10m

*Note: The training duration of 31m 31s on 19th Oct at 14:00 was identified as an outlier and excluded from the mean calculation.

The model training times were recorded for each experimental session, and the results are summarized in Table 1. The average training duration for GGN was approximately 19m 43s, compared to a consistent 10m for the cloud GPU. This indicates a performance differential where GGN took about 97.3% longer to complete the same tasks.

5.2 Cost Analysis

Table 2. Cost (in USD) for GGN and Control

Date	Time	GGN Cost	Control Cost
19th Oct 2024	09:00	$ 0.00512	$ 0.0065
19th Oct 2024	14:00	$ 0.00812	$ 0.0065
19th Oct 2024	00:00	$ 0.0168	$ 0.0065
20th Oct	09:00	$ 0.00601	$ 0.0065
20th Oct	14:00	$ 0.00792	$ 0.0065
20th Oct	00:00	$ 0.0171	$ 0.0065
21st Oct	09:00	$ 0.00484	$ 0.0065
21st Oct	14:00	$ 0.00712	$ 0.0065
21st Oct	00:00	$ 0.0152	$ 0.0065

The cost of using the GGN was evaluated based on the average operational expenses recorded during the experimental period. The costs are summarized in Table 2.

The average cost for GGN usage was approximately $0.00980, compared to $0.0065 for the cloud GPU. While the GGN is more cost-effective during specific periods, its overall average cost is higher due to variability in network performance and operational efficiency.

5.3 Statistical Summary

The statistical summary of the experimental data is provided in Table 3

Table 3. Statistical Summary of Training Duration and Costs

Metric	GGN (Training Duration)	Control (Training Duration)	GGN (Cost)	Control (Cost)
Mean	19m 43s	10m	$ 0.00980	$ 0.0065
Variance	74,970s²	0m	$ 0.0000257	0
Standard Deviation (S.D.)	4m 34s	0m	$ 0.00507	0
95% Confidence Interval	16m 12s - 23m 13s	10m	$0.00590 - $0.01370	$ 0.0065

6 Discussion

The experimental implementation of the Global GPU Network (GGN) reveals several significant findings regarding decentralised GPU computing solutions. The analysis focused on three key aspects: processing performance, temporal variations, and economic viability, each providing unique insights into the system's practical applicability.

The processing time analysis yielded mixed results, with a mean processing time of 19 minutes and 43 seconds (SD = 4:34). This significantly exceeded our hypothesized 20% increase over the 10-minute baseline, with the 95% confidence interval ranging from 16 minutes 12 seconds to 23 minutes 13 seconds. This variance indicates that while the system is functional, it currently falls short of optimal performance expectations, particularly when compared to traditional GPU solutions like the Tesla T4 ($1,532.59).

A notable pattern emerged in the temporal analysis, with distinct performance variations across different time slots. Sessions at 08:00 IST (03:30 UTC) and 00:00 IST (18:30 UTC) demonstrated superior performance, while 14:00 IST (08:30 UTC) sessions consistently showed longer processing durations. This pattern strongly suggests that blockchain network congestion plays a crucial role in system performance, a finding that has significant implications for deployment strategies and user recommendations.

The economic analysis conducted during October 19-21, 2024, revealed promising results. The average smart contract deployment cost remained stable at $0.00980, while delivering significant cost benefits across different time slots. Morning sessions (09:00 IST) proved most economically advantageous, offering 50% discounts to GPU users while maintaining doubled developer revenue. The 14:00 IST sessions provided 40% discounts, and even the 00:00 IST sessions maintained 25% discounts, all while preserving the doubled developer revenue benefit. This consistent economic performance across different time slots suggests a robust economic model, despite varying performance metrics.

Implementation Challenges and Future Prospects The current implementation faces several challenges, primarily centered around processing time stability and performance during high blockchain congestion periods. While the system demonstrates significant economic benefits with consistent 25-50% cost savings during off-peak hours, the performance variability suggests the need for optimization. Future developments should focus on improving network stability during peak congestion periods and implementing more sophisticated load balancing mechanisms.

Several solutions and improvements are under consideration for future iterations. These include implementing dynamic pricing mechanisms that better reflect network conditions, developing advanced node selection algorithms to optimize performance across different time zones, and exploring hybrid models that could combine the cost benefits of decentralized systems with the stability of traditional GPU infrastructure. Additionally, research into blockchain congestion prediction models could help users optimize their deployment timing for maximum cost-effectiveness.

The relationship between blockchain network activity and system performance suggests that future implementations might benefit from a more sophisticated scheduling system that could automatically route tasks to optimal time slots based on historical performance data. This could help maintain the economic benefits while minimizing the impact of performance variations on end-users.

When evaluated holistically, the GGN demonstrates promising potential as a cost-

effective alternative to traditional GPU solutions, particularly for users prioritizing cost savings over consistent processing times. The system's ability to maintain significant cost advantages (25-50% savings) while doubling developer revenue represents a substantial achievement in the field of decentralized computing. However, the performance variability and sensitivity to blockchain network conditions indicate that further development is needed before the system can fully compete with traditional GPU solutions across all use cases.

These findings suggest that while the Global GPU Network has successfully established a viable economic model for decentralized GPU computing, its technical implementation requires refinement to achieve more consistent performance metrics. The path forward likely involves balancing the demonstrated economic benefits with enhanced performance stability, particularly during periods of high network stress.

7 Conclusion

This research investigated the viability of a decentralized GPU computing solution through the implementation and evaluation of the Global GPU Network (GGN). The study addressed the pressing challenges of GPU accessibility and equitable value distribution in the current technological landscape, while examining the potential of mobile computing devices as an alternative computational resource.

While the GGN demonstrated substantial economic benefits, offering 25-50% cost savings compared to traditional cloud GPU services and doubling developer revenue, the system's performance metrics indicated areas requiring optimization. The average processing time of 19 minutes 43 seconds exceeded the hypothesized 20% performance differential, suggesting that technical refinements are necessary for competitive viability.

Notable temporal variations in system performance, particularly during high blockchain congestion periods, highlight the need for more sophisticated load balancing and scheduling mechanisms. These findings contribute to the broader understanding of decentralized computing architectures and their practical implementation challenges.

Future research should focus on optimizing network stability, developing advanced node selection algorithms, and implementing dynamic pricing mechanisms that better reflect network conditions. Despite current limitations, the GGN's demonstrated economic benefits and innovative approach to resource utilization represent a significant step toward democratizing GPU access and creating more equitable value distribution in the computing ecosystem.

Bibliography

Apple unveils all-new iPad Air with A14 Bionic, Apple's most advanced chip. (2020, 15/9/2020). https://www.apple.com/newsroom/2020/09/apple-unveils-all-new-ipad-air-with-a14-bionic-apples-most-advanced-chip/

Ewa, W.-J., Milosz, P., Martyna, K., & Michal, N. (2017). Apple products: A discussion of the product life cycle. 2017 International Conference on Management Science and Management Innovation (MSMI 2017),

Li, T., Xia, T., Wang, H., Tu, Z., Tarkoma, S., Han, Z., & Hui, P. (2022). Smartphone App Usage Analysis: Datasets, Methods, and Applications. *IEEE Communications Surveys & Tutorials*, *24*(2), 937-966. https://doi.org/10.1109/comst.2022.3163176

Lindner, J. (2024). *Key Average Website Traffic Statistics Revealed in Latest Study* (GITNUX Report 2024, Issue. https://gitnux.org/average-website-traffic-statistics/

Maymounkov, P., & Mazières, D. (2002). Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In P. Druschel, F. Kaashoek, & A. Rowstron, *Peer-to-Peer Systems* Berlin, Heidelberg.

Moorhead, P. (2024, 12/9/2024). Antitrust Probes Into Nvidia: What Are The Implications? *Forbes*.

Pierro, G. A., & Rocha, H. (2019, 27-27 May 2019). The Influence Factors on Ethereum Transaction Fees. 2019 IEEE/ACM 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB),

Rogers, A. (2024, 9/9/2024). US accuses Google of dominating ad tech market as antitrust trial begins. *Finical Times*. https://www.ft.com/content/b6225aba-b288-45dd-abdc-00a7bbab6842
Shah, R. (2024). G²N: Global GPU Network. 12.
Strzala, A. (2021). GPU Prices Increased by Up to 300% and May Continue to Rise. Retrieved 11 June 2021, from https://www.gamepressure.com/newsroom/gpu-prices-increased-by-up-to-300-and-may-continue-to-rise/ze33fb
Varatharajan, P. (2024). Google AdSense Updates: Revenue-share Structure and Moving to CPM. https://headerbidding.co/google-adsense-updates-revenue-share-structure-moving-to-cpm/
Robinson, J. (1969). Monopoly Equilibrium. In J. Robinson (Ed.), *The Economics of Imperfect Competition* (pp. 47-59). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-15320-6_4

Authors Contribution
R.S. is the sole author of this manuscript and is responsible for all aspects of the research presented. R.S. conceived the idea of the Global GPU Network (GGN), conducted the performance benchmarking and comparative analysis against traditional GPU computing models, wrote the main manuscript text, and prepared all figures and tables. R.S. also reviewed and revised the manuscript to ensure its accuracy and clarity.

Acknowledgments
The author is grateful to Ahmedabad International School for their support and permission to publish this research.

Availability of data and material
Data booklet and calculations are provided in calculations.tex.

Competing interests
The author declares that they have no competing interests.

Funding
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.