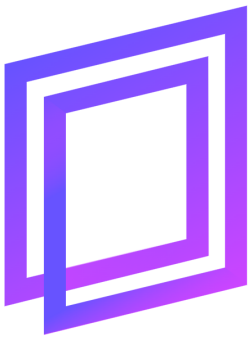
Tokenisation for Decentralised HPC

Creating the First HPC Index Benchmark



DeepSquare

Author
Dr. Florin Dzeladini florin@deepsquare.io

Contributors:
Philippe Thiault philippe@deepsquare.io
Christophe Lillo lillo@deepsquare.io
Christophe Ozcan christophe.ozcan@crypto4all.com
Charly Mancel charly@deepsquare.io

25 April 2023

Contents

Abstract

DeepSquare presents a novel High-Performance Computing as a Service solution designed to meet the growing demand for computing power by businesses. The platform utilises two tokens, GFL and DPS, to standardise computing resources and monetise them for industrial clients and Grid Partners. The GFL token ensures optimal allocation of computing power, while the DPS token allows businesses to access computing resources without owning the infrastructure. The platform's halving mechanism adjusts the token supply, ensuring the sustainability and stability of the ecosystem. Over the first 10 years, Grid Partners are rewarded through inflation, enabling DeepSquare to offer competitive pricing to clients. The platform is designed with environmental sustainability, security, stability, and regulatory compliance in mind, making it an attractive choice for businesses in need of high-performance computing resources.

1 Introduction

1.1 High-Performance Computing (HPC) Demand Overview

The importance of High-Performance Computing (HPC) has grown due to the increased use of data and AI in business. Applications such as Natural Language Processing (NLP), computer vision, and drug discovery require substantial computational power. The demand for HPC has surged as companies seek to derive insights from large data sets and train complex AI models. OpenAI's GPT-3 and deep learning algorithms used in computer vision require vast computing resources to train effectively. HPC is critical for AI-powered systems and expected to continue growing. Cloud computing has made HPC services accessible for companies of all sises, unlike traditional high-performance clusters.

1.2 High-Performance Computing as a Service (HPCaaS)

High-Performance Computing as a Service (HPCaaS) is a cloud-based solution that offers on-demand HPC resources, eliminating the need for costly infrastructure investment and maintenance. High-Performance Computing as a Service (HPCaaS) providers offer a range of services, including technical support, software installation, and data management, allowing users to focus on their core business activities. Compared to traditional HPC models, HPCaaS offers greater flexibility, faster deployment, and lower costs. It enables companies to access computing power that would otherwise be unavailable and reduces carbon footprint by utilising low-carbon data centers.
Microsoft, IBM, and other software providers offer HPCaaS options suitable for various industries and uses. DeepSquare's dual-token approach, combining an accurate measure of computing power allocation with a reward-based compensation scheme, sets it apart in the HPCaaS market.
In summary, HPCaaS offers businesses access to computing power without the expense and complexity of infrastructure ownership and maintenance. It is an attractive option for all businesses, from start-ups to large enterprises. HPCaaS providers like DeepSquare will play an increasingly important role in enabling businesses to leverage the power of data and AI for innovation and growth as HPC demand continues to rise.

2 DeepSquare's HPCaaS Positioning in the Market

DeepSquare offers HPCaaS services with a unique positioning in the market due to its dual-token approach. The objective is to create an HPC index - through the emission of GFL utility tokens - measuring the precise amount of computational resources allocated by DeepSquare providers called "Grid Partners" at any point in time and offer an automatised compensation-based incentive to ensure that the ecosystem remains attractive and sustainable.
DeepSquare's approach involves the tokenisation of distributed computing power through the creation of two dedicated tokens: the GFL utility Token and the DPS payment Token. The GFL Token measures the available computing power in PetaFlop with the rate [1 GFL = 1 PetaFlop] on the DeepSquare network.

The GFL token is allowing real-time auditing and tracking of changes in computing power within DeepSquare HPC decentralised network. The quantity of GFL tokens is determined based on the evolution of available computing power (GPUS, CPUs, etc..) from Grid Partners. The GFL token is a non-tradeable utility token only dedicated to measuring the computing power offered by all providers of the DeepSquare network at any time with a payment every 15 days.
The DPS Token, on the other hand, is a payment token. It represents the average reward for computing power provided by partners in the DeepSquare network, and is used to offer clients preferential access to compute resources. Grid Partners who provide computing power will receive periodic and evolving remuneration paid in DPS tokens based on the amount of GFL tokens representing the affected computer power. The remuneration for providing computing power is based on 20% of the amount of GFL tokens held, paid in DPS tokens at a starting ratio of 1 DPS equal to 100 GFL tokens. This ratio will change over time according to a deflationary mechanism designed to reduce the total supply of DPS tokens, as describe in section 4.

3 GFL Token Technical Overview

3.1 Standardising HPCaaS with GFL

The GFL token, created by DeepSquare, serves as a standardised unit of measurement to quantify the computing power contributed by Grid Partners. This standardisation allows for consistent and transparent comparisons across various platforms and providers, enabling informed decisions in the HPCaaS market.
DeepSquare aims to promote the adoption of GFL through collaboration with industry partners and organisations, emphasising the advantages of GFL over other metrics. GFL provides a consistent and precise measurement of computing power across different platforms and providers, ensuring easy comparison and evaluation of HPCaaS offerings.
Incentivising Grid Partners to maintain or expand their computing power infrastructure, GFL tokens contribute to a stable and reliable supply of computing power for HPCaaS users. Periodic and evolving remuneration in DPS tokens motivates Grid Partners to expand and invest in their infrastructure.

3.2 GFL Emission Strategy

The GFL token is a utility token representing an aggregate of computing power available on the network and it is rebalanced every 15 days to guarantee continuous availability. GFL tokens cannot be listed on exchange platforms; instead, they represent the computing power allocated by each provider within the DeepSquare decentralised network. Think of GFL tokens as analogous to the network hashrate in proof-of-work blockchains, reflecting the amount of computing power dedicated to the network by providers.
Using a Proof-Of-Computing-Power allocation (PoCP) mechanism, the total supply of GFL tokens is dynamic, not bound and rebalanced every 15 calendar days to ensure the incentivisation in DPS tokens based on the real allocation of the computing power. This mechanism ensures that the quantity of underlying index (computing power) remains in equilibrium with the number of GFL tokens distributed among the

network, allowing the ecosystem to maintain a high ratio of occupancy. The tight link between GFL and DPS tokens forms the basis of the grid providers' incentivisation.

3.2.1 Total supply

The total supply of GFL tokens is given by the following formula:

$$\sum_{t=0}^{t+1} \Delta\text{GFL}(t) = f(\text{PCTR})(t+1) - f(\text{PCTR})(t)$$

Where

- ΔGFL is the variation of GFL tokens during a period,
- $f(\text{PCTR})(t)$ is the function that calculates the total computing power allocated within the DeepSquare network at time t

3.2.2 Rebalancing GFL tokens

The GFL token is rebalanced every 15 days using the median computing power allocated by providers on the DeepSquare network during the 15 days period. This rebalancing mechanism ensures that the quantity of underlying indicator (computing power) is maintained in equilibrium with the number of GFL tokens and the real computational power allocated within the network. The rebalancing process, which is automated and based on the computing power provided by Grid Partners over the entire 15-day period, incentivises partners to consistently provision computing power.

3.2.3 Proof-Of-Computing-Power allocation (PoCP)

The Proof-Of-Computing-Power allocation (PoCP) is a mechanism designed to ensure the accurate representation and measurement of the computing power provided by each provider on the DeepSquare network. Unlike traditional blockchain mining systems in the old world of cryptocurrencies, the PoCP focuses on the completion of HPC tasks rather than mining activities.
As a true utility token, the GFL token is not intended to be exchanged on centralised or decentralised digital asset platforms. Instead, it allows for near real-time measurement of the computing power allocated by each provider on the DeepSquare network through the holding of GFL tokens. The GFL tokens serve as an index of the aggregate global computing power.
The PoCP mechanism works by verifying the computing power allocated by each provider on the DeepSquare network, ensuring that the GFL tokens accurately represent the available resources. This process involves the validation of the HPC tasks performed by the grid providers, confirming their contribution to the network's overall computing power.
More detailed information about the Proof of Compute Power (PoCP) mechanism can be found in the dedicated document titled "Proof of Compute Power (PoCP)." This document provides a comprehensive explanation of the PoCP methodology and its implementation within the DeepSquare ecosystem.

In summary, the PoCP mechanism allows DeepSquare to maintain an accurate representation of the computing power available on its network, ensuring that the GFL tokens serve as a reliable and standardised unit of measurement for our decentralised grid of high-performance computing as a service (HPCaaS).

3.3 Example scenario of GFL token adjustment

Let's consider an example scenario to better understand the GFL token adjustment mechanism:
Suppose DeepSquare has initially an available computing power of 100 PetaFlops, equivalent to 100 GFL, from its own infrastructure and partner suppliers. By January 16, 2023 (00:00 UTC), the computing power has decreased by 1 PetaFlop, corresponding to 1 GFL, over the first 15 days.
The Burn & Mint equilibrium mechanism proceeds to readjust the available GFL quantity using the formula:

$$\text{GFL}_{t+1} = f(\text{PCTR})_{t+1} - f(\text{PCTR})_t$$

Where:

- GFL_{t+1} is the total supply of GFL at the end of the adjustment period.
- $f(\text{PCTR})_t$ is the computing power of the DeepSquare network at the end of the adjustment period, expressed in PetaFlops over the 15 days period.

The total supply of GFL is then adjusted to 99 GFL tokens.
This adjustment ensures that the GFL token supply is always in line with the available computing power on the DeepSquare network. As the computing power increases or decreases, the total supply of GFL tokens is adjusted accordingly.
By using the GFL token as a standardised unit of measurement for computing power, DeepSquare can offer its clients preferential access to this computing power for various use cases across different sectors. The DPS token allows businesses with recurring needs for significant computing power, security guarantees, and simulation environments to use the available computing power on the DeepSquare network. The amount of DPS tokens held corresponds to the right to use a specific amount of computing power, based on an initial ratio of 100 GFL equals 1 DPS.

3.4 Provider Selection and Decentralisation

The selection of compute resource providers plays a crucial role in maintaining the quality and reliability of the DeepSquare network. As the network grows and evolves, the provider selection process must adapt to ensure that it remains decentralised and meets the diverse requirements of its users and providers. This section discusses the initial centralised provider selection process, the envisioned transition towards a more decentralised model, and the potential implementation of decentralised governance to involve the community in decision-making related to provider selection.

3.4.1 Initial Centralised Provider Selection

In the initial stages of the DeepSquare network, the selection of compute resource providers will be managed centrally by DeepSquare. This centralised approach ensures that the network is built upon a reliable and high-quality foundation of HPCaaS providers. To achieve this, DeepSquare will evaluate potential providers based on a set of predefined criteria, taking into consideration factors such as the type of hardware, SLA, sustainability, and overall performance.
By establishing a robust centralised provider selection process, DeepSquare aims to create a strong and dependable grid of HPCaaS providers. This will enable the DeepSquare network to deliver consistent and efficient computing power to its users and form a solid base for future growth and decentralisation.

3.4.2 Transitioning to Decentralised Provider Selection

As the DeepSquare network matures and expands, the goal is to transition from a centralised provider selection process to a more decentralised model. In this approach, the responsibility of selecting compute resource providers would gradually shift from DeepSquare to the network's stakeholders, including token holders, Grid Partners, and HPCaaS users.
To achieve this transition, a decentralised governance system will be implemented to allow stakeholders to participate in the decision-making process related to provider selection. This system may involve voting mechanisms, where token holders can vote on proposals related to the admission of new providers or the adjustment of criteria for provider selection. The decentralised governance system could also include a reputation system, where providers are evaluated and ranked based on their performance, hardware, SLA, sustainability, and other relevant factors.

3.4.3 Decentralised Governance and Community Involvement

The decentralised governance system plays a crucial role in involving the community in the provider selection process. It ensures that the DeepSquare network remains adaptable, responsive to user needs, and maintains a high standard of service.
This system could involve mechanisms such as voting, proposals, and stake-based influence, allowing token holders to have a say in matters related to provider selection, network upgrades, and other important decisions. By incorporating the input of the network's stakeholders, DeepSquare aims to create a balanced ecosystem that can effectively address various constraints and requirements, such as hardware type, SLA, sustainability, and performance.

3.4.4 Summary

In summary, the transition towards a decentralised provider selection model is an important step in fostering a more inclusive and robust network. It ensures that the DeepSquare network remains adaptable to changing needs and continues to deliver a reliable and efficient HPCaaS ecosystem.

4 Monetising computing power with the DPS token

DeepSquare understands the importance of providing incentives for businesses to hold and utilise DPS tokens, as well as motivating Grid Partners to contribute computing resources to the network. Two primary benefits arise from holding DPS tokens for businesses:

1. **Halving Process:** Over time, the halving process increases the value of the DPS token. As the number of DPS tokens associated with GFL decreases, the value of the remaining DPS tokens increases, benefiting both businesses that hold them and Grid Partners providing computing power. This process is similar to the Bitcoin mining, where the Halving event in Bitcoin's history have had a significant impact on the price of the cryptocurrency, particularly in its early days. Investors and traders anticipated the reduced supply of new Bitcoins that would enter the market after each halving event, leading to an increase in demand and upward pressure on the price. However, it is important to note that as the Bitcoin ecosystem has matured, many other factors have come into play that can influence the price of the cryptocurrency. While the halving can still be a significant factor, it is not a guarantee of future price increases.
2. **Discounts on Compute Power:** Businesses can benefit from discounts when paying for compute power with DPS tokens. By using DPS tokens as the payment method, clients can access the necessary computing resources at a lower cost, providing additional incentives to hold and use DPS tokens within the DeepSquare ecosystem.

These benefits encourage businesses to hold and use DPS tokens, further driving demand for computing power on the DeepSquare network. Meanwhile, Grid Partners are incentivised to contribute and maintain their computing resources on the network, as they receive a periodic and evolving remuneration paid in DPS tokens. By offering an increasing value proposition and discounts for paying with DPS tokens, DeepSquare can ensure that clients have access to the computing power they need while incentivising Grid Partners to maintain or increase the power available on the network.

4.1 GFL-DPS Relationship Model

The relationship between GFL and DPS tokens is an essential aspect of the DeepSquare ecosystem. The initial relationship is defined as:

$$1\text{DPS} = 100\text{GFL}.$$

The emission of DPS tokens is subject to a halving process. This halving process occurs either every year **or** when a certain cumulative amount of GFL is reached, whichever comes first. This ensures that the amount of DPS issued is mathematically bound and helps maintain the value of the DPS token over time. The halving formula can be expressed as follows:

$$\text{DPS}_n = \frac{\text{GFL}_n}{2^n} \times R,$$

where:

- DPS_n is the number of DPS tokens after the n^{th} halving,
- GFL_n is the cumulative amount of GFL at the n^{th} halving,
- n is the number of halvings that have occurred, and
- R is the initial ratio of DPS to GFL (in this case, $R = \frac{1}{100}$).

The halving process helps maintain the balance between the GFL and DPS tokens, ensuring that the value of DPS tokens remains stable over time. As the halving progresses, the relationship between GFL and DPS tokens is adjusted, providing an increasing value proposition for both businesses that hold DPS tokens and Grid Partners that contribute computing resources to the DeepSquare network.
By managing the GFL-DPS relationship through the halving process, DeepSquare can create a sustainable and evolving ecosystem that offers preferential access to computing power for its clients, while also incentivising Grid Partners to maintain or increase the power available on the network.

4.2 Managing network growth through DPS reward pool replenishment

The replenishment of the DPS reward pool for Grid Partners is a continuous process that is always active. This is achieved by allocating DPS tokens based on the amount of computing power sold on the network (i.e. proportional to the amount of GFL tokens). This mechanism effectively makes the DPS tokens available again to compensate for inflation, controlling the rate at which halving occurs and ensuring that the GFL-DPS relationship remains stable and sustainable over a desired time period (see Figure 1). By continually replenishing the DPS reward pool proportionally to the network usage, DeepSquare can maintain a balanced and responsive ecosystem.

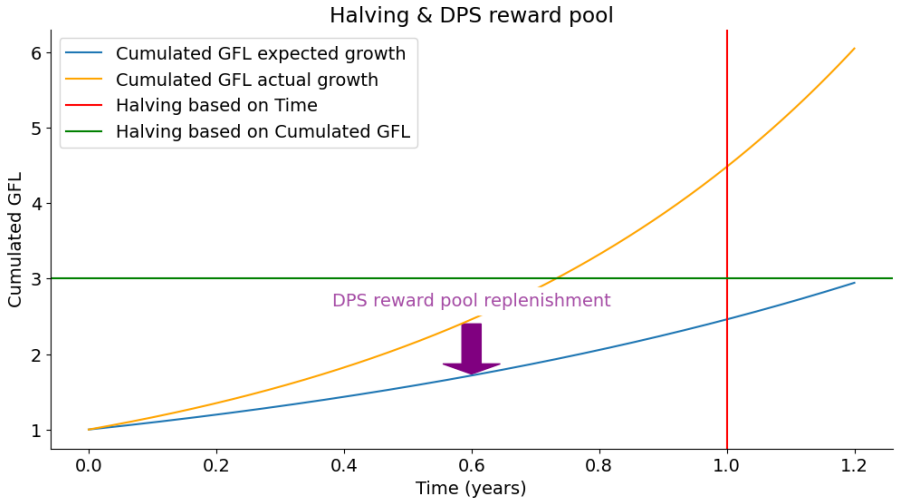


Figure 1: Illustration of the halving process and its impact on the growth of cumulated GFL, showing the effect of DPS pool replenishment

$$\text{GFL}_{new} = \text{GFL}_{pool} - (\text{DPS}_{used} \times R),$$

where:

- GFL_{new} is the new amount of GFL in the pool after the buyback,
- GFL_{pool} is the current amount of GFL in the pool,
- DPS_{used} is the number of DPS tokens used for the buyback, and
- R is the initial ratio of DPS to GFL (in this case, $R = \frac{1}{100}$).

The updated figure illustrates the expected growth of cumulated GFL and the actual growth with the corrective action taken to buy back tokens. This buyback ensures that the halving occurs at the expected time (in this case, at year 1) and not before. By actively managing the halving process and the GFL-DPS relationship, DeepSquare can ensure the long-term stability of its ecosystem and provide optimal access to computing power for its clients and Grid Partners.

5 Addressing Potential Concerns

In this section, we address potential concerns related to security, stability, regulatory compliance, and other aspects of the DeepSquare platform and tokens.

5.1 Security

DeepSquare takes security very seriously, as the platform handles sensitive data and provides critical computational resources to clients. The following measures are implemented to ensure security at all levels:

- End-to-end encryption: Data transmitted between clients and Grid Partners is encrypted using industry-standard algorithms to prevent unauthorised access.
- Secure infrastructure: The DeepSquare platform follows best practices in infrastructure security, including regular updates and patches, network segmentation, and the use of intrusion detection and prevention systems.
- Access control: DeepSquare implements strict access controls, ensuring that only authorised users and Grid Partners can access the platform and resources.

5.2 Stability

Maintaining a stable and reliable platform is essential for DeepSquare's clients and partners. The platform is designed to handle high workloads, ensuring smooth operation even during peak demand periods. DeepSquare also employs a robust monitoring system to detect and resolve any issues that might arise promptly.

5.3 Regulatory Compliance

DeepSquare is committed to comply with all applicable regulations, including those related to data protection, privacy, and the handling of digital assets. DeepSquare is actively monitoring the regulatory landscape to ensure ongoing compliance and will adapt its policies and practices accordingly.

5.4 Token Economics

DeepSquare has designed the GFL and DPS tokens with a long-term vision, incorporating mechanisms such as halving and DPS buybacks to ensure the sustainability of the token economy. These measures help maintain a healthy balance between the supply and demand for tokens while providing incentives for Grid Partners to continue supporting the platform.

5.5 Interoperability

To ensure that the DeepSquare platform remains flexible and adaptable to future developments in the HPC industry, the platform is built with interoperability in mind. DeepSquare supports a wide range of HPC technologies and is designed to integrate seamlessly with other platforms and systems.
In conclusion, DeepSquare is committed to addressing potential concerns and providing a secure, stable, and compliant platform that benefits both clients and Grid Partners. By continuously monitoring and adapting to the evolving HPC and regulatory landscapes, DeepSquare aims to remain at the forefront of the HPCaaS market.

5.6 Environmental Impact

DeepSquare is committed to promoting sustainability and minimising the environmental impact of its platform. This commitment is reflected in the design of the GFL token and the platform's features.

5.6.1 Incorporating Environmental Factors into GFL Calculation

The GFL calculation takes into account several factors that directly influence the environmental footprint of the HPC resources provided by Grid Partners. These factors include:

- Power efficiency: DeepSquare encourages Grid Partners to use energy-efficient hardware to reduce power consumption. The GFL calculation can integrate power consumption metrics to reward partners that operate with lower energy consumption, thus promoting more sustainable practices.
- Cooling technologies: Efficient management of thermal constraints can significantly impact energy consumption and the overall environmental footprint. The GFL calculation can take into account metrics related to thermal management, such as cooling efficiency, to incentivise Grid Partners to optimise their HPC infrastructure accordingly.
- Software optimisations: DeepSquare recognises the importance of software optimisations in improving the energy efficiency of HPC workloads. The platform encourages the use of optimised software and can integrate metrics related to software performance into the GFL calculation, rewarding partners that contribute to reducing the overall energy consumption of the platform.

By incorporating these environmental factors into the GFL calculation, DeepSquare ensures that its platform rewards and incentivises sustainable practices among Grid Partners.

5.6.2 Promoting Renewable Energy Sources

In addition to the measures mentioned above, DeepSquare actively encourages the use of renewable energy sources among its Grid Partners. By supporting Grid Partners that utilise clean energy sources such as solar, wind, or hydroelectric power, DeepSquare aims to reduce the platform's overall carbon footprint and contribute to a more sustainable HPCaaS ecosystem.
In summary, DeepSquare is dedicated to promoting sustainability and minimising the environmental impact of its platform. By incorporating environmental factors into the GFL calculation and encouraging the use of renewable energy sources, DeepSquare aims to create a more sustainable and environmentally conscious HPCaaS platform.

6 Adaptable Computational Power Index

In High-Performance Computing as a Service (HPCaaS) platforms like DeepSquare, establishing a fair and adaptable system for measuring computational power is essential for efficient resource allocation and payment distribution. Traditional metrics, such as FLOPS, may not accurately represent accelerators performance for diverse workloads. Thus, a more comprehensive index is needed.
In this section, we introduce a methodology for constructing an adaptable computational power index, accounting for factors such as memory bandwidth, capacity, and hardware features like GPU Compute Capabilities, CPUs instruction sets or networking capabilities such as RDMA. This index, represented by the GFL token, allows for a more accurate performance evaluation and fair payment distribution among users contributing resources to the grid. Furthermore, the GFL token-based index can be refined and adapted to cater to evolving requirements in the HPCaaS ecosystem.

6.1 Building the Index

To create the computational power index, we follow a series of steps that involve calculating the average FLOPS, identifying workload categories, evaluating performance factors, and applying workload-specific adjustments. The steps are as follows:

1. **Determine the average FLOPS**: Calculate the average FLOPS for the GPUs to be included in the index. This will serve as a baseline performance metric. One can use vendor-reported FLOPS or benchmarks to calculate the average.
2. **Identify workload categories**: Classify the desired workloads to support on the grid, such as machine learning training, inference, and rendering tasks.
3. **Evaluate performance factors for each workload**: For each workload category, key factors that impact accelerators performance are identified. These factors can include memory bandwidth, memory capacity, parallelism, and specific hardware features like Tensor Cores or ray-tracing capabilities.
4. **Calculate workload-specific adjustments**: For each workload category, a weighting factor or adjustment ratio is determined based on the performance

factors identified in step 3. Benchmarks results and empirical data are used to estimate these adjustments.

5. **Apply adjustments to the average FLOPS**: The average FLOPS is multiplied by the workload-specific adjustment ratios to obtain a corrected FLOPS value for each workload category. This will allow the adaptation of the ratio and create a more meaningful index for the grid.
6. **Payment calculation**: Use the corrected FLOPS values to calculate the payment for providers based on the actual performance of their hardware for a specific workload. Here one can either use the corrected FLOPS as a direct multiplier or establish tiers based on the corrected FLOPS to create different payment levels.

This step-by-step approach allows the creation of a flexible and adaptable index that accounts for the performance requirements of different workloads. By adjusting the FLOPS based on workload-specific factors, the index ensures that payments are meaningful and reflective of the actual performance of the hardware.

6.2 Mathematical Formulation

The following subsections describe the formulas and calculations required to implement this methodology.

$$FLOPS_{corrected,i} = FLOPS_{avg} \times W_i \tag{1}$$

Where:

- $FLOPS_{corrected,i}$: The corrected FLOPS for workload i
- $FLOPS_{avg}$: The average FLOPS for the considered accelerators
- W_i: The adjustment ratio for workload i

The adjustment ratio W_i for each workload is calculated using a weighted sum of normalised performance factors relevant to that specific workload:

$$W_i = \alpha \times P_{1,i} + \beta \times P_{2,i} + \gamma \times P_{3,i} + \cdots + \omega \times P_{n,i} \tag{2}$$

Where:

- W_i: The adjustment ratio for workload i
- $P_{k,i}$: The normalised performance factor k for workload i (e.g., memory bandwidth, memory capacity, hardware features)
- $\alpha, \beta, \gamma, \cdots, \omega$: The weights assigned to each performance factor (summing up to 1), representing the relative importance of each factor in the performance of workload i

Finally, the payment for each workload can be calculated using the corrected FLOPS as follows:

$$P_i = C \times FLOPS_{corrected,i} \times T \tag{3}$$

Where:

- P_i: The payment for workload i
- C: The cost per FLOPS per unit of time (e.g., per 15 days)
- $FLOPS_{corrected,i}$: The corrected FLOPS for workload i
- T: The duration of the workload in the chosen unit of time (e.g., 15 days)

6.3 Implementing the Computational Power Index

In this section, we discuss the process of implementing the computational power index based on the mathematical formulation presented earlier. We will cover the necessary steps to prepare and organise the performance data, compute workload-specific adjustments, calculate payments for resource providers, and ensure the adaptability of the index to accommodate technological advancements and changing workload requirements. By following this implementation process, the computational power index can provide a fair and accurate representation of the actual performance of various hardware components for different workloads.

6.3.1 Data Preparation

To implement the mathematical formulation, the first step is preparing the required data. This involves collecting and organising performance data for various computing components, normalising performance factors, and establishing workload categories.
Collecting and organising performance data requires compiling a list of popular and representative models for each type of computing component, such as GPUs and CPUs. Gather detailed performance data from manufacturers' websites, technical datasheets, and reputable benchmark databases.
Normalisation of performance factors involves converting raw performance metrics into a common scale, allowing for easy comparison between different computing components. The normalisation process typically consists of dividing each performance metric by the maximum value observed within the dataset or rescaling the values to a range between 0 and 1.
Establishing workload categories involves classifying the desired workloads to support on the grid, such as machine learning training, inference, and rendering tasks. For each workload category, determine the critical performance factors affecting accelerators performance, including memory bandwidth, memory capacity, parallelism, and specialized hardware features, such as Tensor Cores or ray-tracing capabilities.

6.3.2 Computing Workload-Specific Adjustments

To implement the mathematical formulation, workload-specific adjustments must be computed. This involves calculating adjustment ratios as well as integrating benchmark results and empirical data.
First, calculate the adjustment ratios for each workload category. Use the weighted sum of normalised performance factors relevant to that specific workload, as described in the mathematical formulation. Assign appropriate weights to each

performance factor, considering the relative importance of each factor in the performance of the workload.
Next, integrate benchmark results and empirical data to estimate the workload-specific adjustments. Collect benchmark data for various computing components, including GPUs, CPUs, and other accelerators like FPGAs. Use reputable benchmark databases and widely accepted benchmarking tools and software, such as SPEC, MLPerf, and 3DMark. Analyze the benchmark data to identify performance trends, correlations, and patterns that can inform the adjustment ratios.
Finally, combine the calculated adjustment ratios with the benchmark and empirical data to compute workload-specific adjustments. These adjustments will be applied to the average FLOPS to obtain corrected FLOPS values for each workload category, allowing for a more accurate representation of the GPU's performance for a specific workload and fairer distribution of payments to users who contribute their computing resources to the grid.

6.3.3 Payment Calculation

Implementing the payment calculation requires applying the mathematical formulation established in section 6.3. By utilising the corrected FLOPS values, we can calculate the payment for providers based on the actual performance of their hardware for a specific workload. The payment formula, as shown in Equation 3, combines the cost per FLOPS per unit of time, the corrected FLOPS for the particular workload, and the duration of the workload in the chosen unit of time.
To establish a fair and efficient payment system, it is essential to determine C, which represents the appropriate cost per FLOPS per unit of time that reflects the market value of the computing resources. This cost factor can be determined through market analysis, consultation with industry experts, and by considering operational costs, such as electricity and maintenance.

6.3.4 Iterative Improvement and Adaptation

To ensure the computational power index remains relevant and accurate over time, it is essential to continuously improve and adapt the index as new technologies emerge and workload requirements change. The following steps outline the process of iterative improvement and adaptation:

1. **Periodic Data Updates**: Regularly update the performance data for various computing components, such as GPUs, CPUs, and other accelerators. This includes incorporating new models, updating performance metrics, and revising benchmarks as needed. Regular data updates help maintain the accuracy of the index and account for the latest advancements in technology.
2. **Incorporating New Technologies**: As new hardware technologies become available, evaluate their impact on the computational power index and integrate them into the calculations as necessary. For example, emerging accelerators like AI-specific ASICs or quantum processors may require modifications to the index to account for their unique performance characteristics.

3. **Updating Workload Categories**: Review and update the workload categories periodically to ensure they remain relevant and reflective of the current demands in the market. As new types of workloads emerge or existing ones evolve, it may be necessary to adjust the categories and corresponding performance factors accordingly.
4. **Revising Adjustment Ratios and Weights**: As workload requirements and hardware performance change over time, revisit the adjustment ratios and weights assigned to various performance factors. Update these values based on the latest benchmarks, empirical data, and expert insights to maintain the fairness and accuracy of the index.
5. **Adapting the Payment Model**: Periodically review the payment model to ensure it continues to provide a fair and meaningful compensation for resource providers. This may involve adjusting the cost per FLOPS, updating the payment tiers, or incorporating new factors that influence the value of contributed computing resources.

By following these steps, the computational power index can stay up-to-date with the latest technological advancements and workload requirements, ensuring that it remains a reliable and adaptable tool for quantifying the performance of different accelerators and calculating payments for resource providers on a grid.

6.4 Summary

In this section, we have presented a methodology for creating an adaptable computational power index that accurately represents the performance of various computing resources for different workloads. We have outlined the steps to build the index, the mathematical formulation behind it, and the process of implementing the mathematical formulation. By adjusting the FLOPS based on workload-specific performance factors, the index ensures that payments are more meaningful and reflective of the actual performance of the hardware. The approach is designed to be flexible and adaptable, allowing for continuous improvement and incorporation of new technologies and workload requirements.

7 Conclusion and Outlook

DeepSquare's innovative approach to High-Performance Computing as a Service (HPCaaS) aims to provide businesses with a cost-effective, reliable, and scalable solution for their computing needs. By leveraging the GFL and DPS tokens, DeepSquare has created an ecosystem that benefits both industrial clients in search of computing power and Grid Partners allocating IT resources to the platform. The GFL, which represents the adaptable computational power index, offers a flexible and accurate representation of the performance of various computing resources for different workloads, ensuring that clients receive the most suitable resources for their specific needs.
During the first 10 years of the project, Grid Partners are rewarded through inflation, allowing DeepSquare to offer very competitive pricing for businesses seeking access to high-performance computing resources. This pricing advantage, combined with

the fair and meaningful payments enabled by the computational power index, makes DeepSquare an attractive option for companies looking to optimise their computing infrastructure without incurring significant costs.

As the project matures, DeepSquare will continue to focus on enhancing the platform's features, including refining and adapting the computational power index to accommodate advances in technology, new hardware releases, and changes in workload requirements. In addition, DeepSquare is committed to improving the environmental sustainability of its operations and addressing potential concerns related to security, stability, and regulatory compliance. By doing so, DeepSquare aims to establish itself as a leading player in the HPCaaS market and become a preferred choice for businesses looking to leverage high-performance computing resources.

Looking ahead, DeepSquare is dedicated to further developing and refining its platform to meet the evolving needs of its clients and the broader HPCaaS market. Through ongoing innovation and a focus on fostering a sustainable, secure, and efficient ecosystem, DeepSquare seeks to empower businesses with the computing power they need to drive growth and success in an increasingly competitive and data-driven world.

Glossary

DPS A payment token that ensures payment to Grid Partners. Its tokenomics is tightly linked to the GFL tokens through the GFL/DPS emission relationship. The DPS token is no longer considered a security token and is granted to providers as a return for burning GFL tokens. DPS holders will also receive shares of DeepLabs company based on different rules..

GFL Stands for Guaranteed FLOPS Token. Is a pure utility token representing the disposable computing power on the grid. Providers are paid DPS by burning GFL tokens based on a fixed ratio. GFL tokens cannot be transferred and its supply can theoretically increase without a limit..

Grid Partners Organisation providing IT infrastructure to the DeepSquare computing ecosystem..

Halving event refer to a specific type of event that occurs in the mining process for certain cryptocurrencies, including Bitcoin and several others. During the mining process, new units of a cryptocurrency are created as a reward for miners who solve complex cryptographic puzzles that verify transactions on the blockchain network..

HPC High-Performance Computing.

HPCaaS High-Performance Computing as a Service.

NLP Natural Language Processing.

PoCP Proof-Of-Computing-Power allocation.