cfbenchmarks
a PAYWARD company

CF H100 Neo-Cloud GPU Compute Index

Version: 1.0 28th July 2026

Contents

1 Version History

Version	Date Issued	Summary of Change	Owner
V1.0	28 Jul 2026	Document creation: initial methodology for the CF H100 Neo-Cloud GPU Compute Index	CF Benchmarks Product

2 Overview

Access to accelerator compute capacity, and to NVIDIA H100 GPUs in particular, has become one of the most economically significant inputs to the artificial intelligence ecosystem. Companies training and serving AI models increasingly plan and budget in terms of compute cost rather than headcount or general infrastructure spend alone. Despite this, no single, transparent, rules-based reference price for renting this capacity is widely established. The CF H100 Neo-Cloud GPU Compute Index is intended to serve as a transparent and representative indicator of the on-demand market price of renting a single NVIDIA H100 SXM GPU, expressed in U.S. Dollars per GPU-hour. It is designed to be a rules-based, reproducible benchmark suitable for use as a reference rate, and, subject to further development, as an input to tradable financial products such as perpetual contracts referencing the price of compute.

Input Data

The Index is calculated from machine-readable on-demand rental prices retrieved from selected APIs approved by the Administrator. These comprise direct Provider APIs, as well as marketplace and platform GPU rental APIs through which multiple underlying Providers publish rental prices. A single Provider may accordingly contribute multiple Quotes at each Calculation Time, corresponding to distinct rentable configurations, regions, or machines. Hyperscaler cloud platforms are excluded from the Provider Panel.

No Time-Series Smoothing

The Index does not apply any rolling-window or multi-day smoothing. Each Calculation Time's published value is calculated from that Calculation Time's Eligible Quotes only. This is a deliberate design choice intended to preserve replicability and tradability of the Index for use as a reference rate for derivative products; see the contingency rules in Section 5 for how transient bad data is instead controlled for without smoothing.

3 Definitions

API: Application Programming Interface.

Reference Configuration: A single NVIDIA H100 SXM GPU, offered on an on-demand (as opposed to spot or reserved/committed) basis, with pricing normalized to a single-GPU unit in U.S. Dollars per GPU-hour.

Provider: An accelerator cloud, GPU rental service, or GPU marketplace approved by the Administrator as a pricing source for the Index, whether observed via its own API or via an approved marketplace or platform API. The Provider Panel is defined in Section 6.

Provider Panel: The full set of Providers currently approved for the Index.

Quote: An individual on-demand rental price for the Reference Configuration published by a Provider and retrieved from an approved API, normalized to USD per GPU-hour on a single-GPU basis. A Provider may contribute multiple Quotes at a given Calculation Time (distinct configurations, regions, or machines).

Calculation Time (t): Each time at which the CF H100 Neo-Cloud GPU Compute Index is calculated and published, at the frequency defined in Section 6.

Last Retrieval Time: The most recent time at which a given Quote was successfully retrieved from its source API.

Retrieval Lag Threshold: The maximum permitted period between a Quote's Last Retrieval Time and the Calculation Time. Within this period a Quote is carried forward at its last retrieved value; beyond it, the Quote is disregarded as stale (see Section 5.1).

Provider Aggregate: The arithmetic mean of a Provider's Quotes at a given Calculation Time. The Provider Aggregate is used solely for the contingency screening rules of Section 5; it does not enter the Index calculation directly.

Eligible Provider: A Provider whose Provider Aggregate has passed all of the automated screening rules in Section 5 as of a given Calculation Time.

Eligible Panel (E_t): The subset of the Provider Panel that is Eligible at Calculation Time t.

Eligible Quote: A Quote of an Eligible Provider that has passed the data preparation rules of Section 5.2.

Panel Median: The median of the Provider Aggregates of the Providers surviving the screening rules of Sections 5.1 to 5.3 (that is, after the Outlier Gate) at a given Calculation Time (see Section 4). The Panel Median serves as the cross-sectional reference for the Potentially Erroneous Data check and for reinstatement; it is not itself the published Index value.

Provider Weight Cap (C): The maximum share of total weight that any single Provider's Quotes may carry in the Index calculation at a given Calculation Time (see Sections 4 and 6).

Minimum Panel Size: The minimum number of Eligible Providers required for the Index to be calculated from current Quotes at a given Calculation Time; if it is not met, the previously published Index value is republished (see Sections 5.5 and 6).

Outlier Gate: The contingency rule under which a Provider is temporarily excluded from the calculation if its Provider Aggregate moves by more than a threshold percentage relative to its own previous accepted print (see Section 5.3).

Potentially Erroneous Data Parameter (PEDP): The maximum permitted absolute percentage deviation of a Provider Aggregate from the Panel Median before that Provider is flagged as potentially erroneous (see Section 5.4).

4 Methodology

4.1 Qualitative Description

At each Calculation Time t:

1. All Quotes are retrieved from the approved APIs. Quotes are normalized to the Reference Configuration (USD per GPU-hour, single-GPU basis), duplicate listings of the same underlying capacity are removed, and Quotes not retrieved within the Retrieval Lag Threshold are disregarded; Quotes retrieved within the Retrieval Lag Threshold but not at time t itself are carried forward at their last retrieved value (Sections 5.1 and 5.2).
2. For each Provider p, the Provider Aggregate $a_{p,t}$ is calculated as the arithmetic mean of that Provider's Quotes.
3. Each Provider Aggregate is subjected to the automated screening rules of Section 5, in sequence: first the Outlier Gate (Section 5.3), then the Potentially Erroneous Data check (Section 5.4), whose Panel Median is calculated over the Providers surviving the Outlier Gate. Providers that fail any check are excluded from the calculation at time t, together with all of their Quotes, and are not part of the Eligible Panel E_t. Excluded Providers are automatically reinstated as set out in Section 5.
4. If the Eligible Panel E_t contains fewer Providers than the Minimum Panel Size (Section 6), a Calculation Failure is recorded and the previously published Index value is republished for that Calculation Time (Section 5.5).
5. Otherwise, all Quotes of the Eligible Panel are assigned equal weight, subject to the Provider Weight Cap: any Provider whose Quotes would carry more than C of the total weight is capped at exactly C, and the excess weight is redistributed pro-rata across the remaining, uncapped Providers in a single pass (no iteration).
6. The CF H100 Neo-Cloud GPU Compute Index at time t is published as the weighted mean of the Eligible Quotes under these weights, with no further smoothing applied.

4.2 Mathematical Representation

The following table shows the symbols used in the mathematical representation of the CF H100 Neo-Cloud GPU Compute Index.

Symbol	Name	Description	Type
t	Calculation Time	The time at which the Index is calculated	Parameter, see Section 6

P	Provider Panel	The full set of approved Providers	Parameter, see Section 6
$Q_{p,t}$	Quote Set	The set of Provider p's Quotes at time t, after the data preparation rules of Sections 5.1–5.2	Input
$q_{i,t}$	Quote	The i-th Quote at time t, normalized to USD per GPU-hour, single-GPU basis	Input
$n_{p,t}$	Quote Count	The number of Provider p's Quotes at time t, $n_{p,t} = \|Q_{p,t}\|$	Calculated
$a_{p,t}$	Provider Aggregate	The arithmetic mean of Provider p's Quotes at time t	Calculated
RLT	Retrieval Lag Threshold	Maximum permitted period between a Quote's Last Retrieval Time and t	Parameter, see Section 6
X	Outlier Gate Threshold	Maximum permitted percentage change of $a_{p,t}$ relative to Provider p's previous accepted print	Parameter, see Section 6
$PEDP$	Potentially Erroneous Data Parameter	Maximum permitted absolute percentage deviation of $a_{p,t}$ from m_t	Parameter, see Section 6
S_t	Post-Gate Panel	The subset of P surviving Sections 5.1–5.3 (staleness, erroneous data, and the Outlier Gate) at time t	Calculated
E_t	Eligible Panel	The subset of S_t that also passes the Potentially Erroneous Data check	Calculated
m_t	Panel Median	The median of $\{a_{p,t} : p \in S_t\}$	Calculated
N_t	Eligible Quote Count	Total number of Eligible Quotes, $N_t = \sum_{p \in E_t} n_{p,t}$	Calculated
C	Provider Weight Cap	Maximum share of total weight carried by any single Provider	Parameter, see Section 6
$w_{i,t}$	Quote Weight	The weight assigned to Quote i at time t after application of the Provider Weight Cap	Calculated

MPS	Minimum Panel Size	Minimum $\lvert E_t \rvert$ required for publication	Parameter, see Section 6
GPU_t	CF H100 Neo-Cloud GPU Compute Index	The Index value published at time t	Calculated

The Provider Aggregate of Provider p at Calculation Time t is:

$$a_{p,t} = \frac{1}{n_{p,t}} \sum_{i \in Q_{p,t}} q_{i,t} \tag{1}$$

The Panel Median at Calculation Time t is calculated over the Post-Gate Panel S_t, i.e. the Providers surviving the staleness, erroneous-data, and Outlier Gate screenings of Sections 5.1–5.3:

$$m_t = \text{MEDIAN}_{p \in S_t}(a_{p,t}) \tag{2}$$

The Potentially Erroneous Data check of Section 5.4 is then applied against m_t; the Eligible Panel E_t comprises the Providers of S_t that also pass that check. Each Eligible Quote initially carries equal weight, so that Provider p's share of total weight is $n_{p,t}/N_t$. The Provider Weight Cap is then applied in a single pass. Let

$$O_t = \{\, p \in E_t : n_{p,t} > C\, N_t \,\}, \qquad U_t = E_t \setminus O_t,$$

denote the over-cap and under-cap Providers respectively ($\setminus$ denotes the set difference: U_t contains the Eligible Providers that are not in O_t), and let the excess weight be

$$\Delta_t = \sum_{p \in O_t} (n_{p,t} - C\, N_t).$$

Provided O_t is non-empty and $\sum_{u \in U_t} n_{u,t} > 0$, the weight of Quote i belonging to Provider p is:

$$w_{i,t} = \begin{cases} \dfrac{C\, N_t}{n_{p,t}} & p \in O_t \\ 1 + \dfrac{\Delta_t}{\sum_{u \in U_t} n_{u,t}} & p \in U_t \end{cases} \tag{3}$$

so that each over-cap Provider carries exactly the share C and the excess is redistributed across the under-cap Providers pro-rata to their Quote counts; otherwise $w_{i,t} = 1$ for all Quotes. The cap is applied in a single pass with no subsequent re-check, and total weight is conserved ($\sum_i w_{i,t} = N_t$). Subject to the Minimum Panel

Size condition of Section 5.5, the CF H100 Neo-Cloud GPU Compute Index at time t is:

$$GPU_t = \begin{cases} \dfrac{\sum_i w_{i,t}\, q_{i,t}}{\sum_i w_{i,t}} & \text{if } |E_t| \geq MPS \\ GPU_{t-1} & \text{otherwise (Calculation Failure, Section 5.5)} \end{cases} \tag{4}$$

5 Contingency Calculation Rules

5.1 Stale Data and Quote Carry-Forward

Each Quote is timestamped with its Last Retrieval Time. A Quote that has been successfully retrieved within the Retrieval Lag Threshold RLT preceding the Calculation Time, but not at the Calculation Time itself, is carried forward at its last retrieved value. If the period between a Quote's Last Retrieval Time and the Calculation Time exceeds RLT, that Quote is disregarded from the calculation of the CF H100 Neo-Cloud GPU Compute Index for that Calculation Time as stale. If this condition is met for all Quotes of all Providers in the Provider Panel at a given Calculation Time, a Calculation Failure occurs (see Section 5.5).

5.2 Erroneous Data

All Quotes are subject to an automated screening for erroneous data:

1. If a Quote is non-numeric, non-positive, or cannot be parsed into the Reference Configuration's unit (USD per GPU-hour, single-GPU basis), it is flagged as erroneous.
2. Where a Provider publishes prices for multi-GPU configurations, each is first normalized to a single-GPU, single-hour basis before being treated as a Quote; if normalization is not possible, the Quote is flagged as erroneous.
3. Duplicate listings of the same underlying capacity (for example, the same machine listed multiple times via a marketplace, or the same Provider configuration observed via more than one API) are de-duplicated, retaining a single Quote. For GPU marketplace sources, only listings from machines verified by the marketplace operator are included.

Quotes flagged as erroneous for a given Calculation Time are disregarded from the calculation of the CF H100 Neo-Cloud GPU Compute Index for that Calculation Time.

5.3 Outlier Gate

All Provider Aggregates are subject to an automated screening for excessive period-on-period change:

1. For each Provider, the absolute percentage change between $a_{p,t}$ and the Provider's own previous accepted print is calculated.
2. If this change exceeds the Outlier Gate Threshold X (Section 6), the Provider is

excluded from the calculation of the CF H100 Neo-Cloud GPU Compute Index at time t, together with all of its Quotes, and the Index is calculated using the remaining Eligible Panel.

3. A Provider excluded under this rule is re-included in the calculation at a subsequent Calculation Time once its Provider Aggregate's absolute percentage deviation from the Panel Median of the remaining Eligible Panel falls within the Reinstatement Threshold applied to the Potentially Erroneous Data Parameter (see Section 5.4 and Section 6). At that point it is reinstated for that Calculation Time and all subsequent Calculation Times, unless it is excluded again under this or any other rule in this Section.

This rule is intended to isolate transient, single-print pricing errors (for example, a temporarily incorrect published rate from a single Provider) without materially distorting the published Index for that Calculation Time, and without requiring any multi-day smoothing of the series as a whole.

5.4 Potentially Erroneous Data

Following the Outlier Gate of Section 5.3, all Provider Aggregates of the Post-Gate Panel S_t are subject to a cross-sectional screening for potentially erroneous data:

1. For each Provider, the absolute percentage deviation of $a_{p,t}$ from the Panel Median m_t is calculated. m_t is calculated over the Providers that survive the screenings of Sections 5.1 to 5.3 — that is, after the Outlier Gate has been applied (Equation 2).
2. If this deviation exceeds the Potentially Erroneous Data Parameter $PEDP$ (Section 6), the Provider is flagged as potentially erroneous and excluded, together with all of its Quotes, from the calculation of the CF H100 Neo-Cloud GPU Compute Index for that Calculation Time.
3. A Provider so excluded remains excluded until its deviation, recalculated at each subsequent Calculation Time, falls within the Reinstatement Threshold (Section 6) applied to $PEDP$, at which point it is reinstated.

If all Providers in the Provider Panel are flagged as potentially erroneous at a given Calculation Time, a Calculation Failure occurs (see Section 5.5).

5.5 Calculation Failure and Minimum Panel Size

If, after applying Sections 5.1 to 5.4, the Eligible Panel E_t contains fewer Providers than the Minimum Panel Size MPS (Section 6), a Calculation Failure is deemed to

have occurred and the CF H100 Neo-Cloud GPU Compute Index is not calculated from Quotes for that Calculation Time; instead, the previously published Index value is republished (Equation 4). The occurrence of any Calculation Failure is reported to the Oversight Function and communicated to licensees via Statuspage.

5.6 Expert Judgement

The Administrator does not utilise expert judgement in the day-to-day calculation of the CF H100 Neo-Cloud GPU Compute Index. In extraordinary circumstances, Expert Judgement may be exercised by the Administrator in accordance with its codified policies and processes, which are available upon request.

6 Index Parameters

The CF H100 Neo-Cloud GPU Compute Index is denominated in U.S. Dollars per GPU-hour for the Reference Configuration.

Name	**CF H100 Neo-Cloud GPU Compute Index**
Ticker	[TBD]
Reference Configuration	NVIDIA H100 SXM, on-demand, USD per GPU-hour, single-GPU basis
Data Sources	Selected APIs approved by the Administrator: direct Provider APIs, and marketplace / platform GPU rental APIs (marketplace listings from verified machines only). Hyperscaler cloud platforms are excluded.
Provider Panel (P)	RunPod; Vast.ai; Denvr, DigitalOcean, Horizon, Hyperstack, IMWT, Lambda, Latitude, Massed Compute, Nebius, Paperspace, Scaleway, Verda, and Voltage Park, via Shadeform. The panel is maintained by the Administrator and reviewed periodically.
Aggregation Function	Equal-weighted mean of Eligible Quotes, subject to the Provider Weight Cap (Equations 3–4)
Provider Weight Cap (C)	25%, applied in a single pass with pro-rata redistribution
Smoothing	None
Calculation & Publication Frequency	Hourly, 24/7, 365 days a year
Retrieval Lag Threshold (RLT)	24 hours (carry-forward within the threshold)
Outlier Gate Threshold (X)	33%

Potentially Erroneous Data Parameter ($PEDP$)	60%
Reinstatement Threshold	50% of $PEDP$, i.e. 30%
Minimum Panel Size (MPS)	5
Decimals Precision	4

7 Methodology and Review Changes

This methodology is subject to internal review by the Administrator and the Oversight Function at least annually. Any changes to this methodology are overseen by the Oversight Function, and in accordance with UK BMR Article 13. All material changes to the methodology shall only be implemented after a consultation process with users and relevant stakeholders that shall be conducted according to the Administrator's policies and overseen by the Oversight Function. Should the Administrator deem it necessary to cease providing the CF H100 Neo-Cloud GPU Compute Index it shall only do so after a consultation process with users and relevant stakeholders that shall be conducted according to the Administrator's policies and overseen by the Oversight Function.

Contact Information

CF Benchmarks Ltd	
Address	**Contact**
CF Benchmarks Ltd 6th Floor One London Wall London EC2Y 5EB	Web: https://www.cfbenchmarks.com Phone: +44 20 8164 9900 Email: contact@cfbenchmarks.com
Formal complaints or concerns regarding CF Benchmarks pricing products must be submitted by Email: complaints@cfbenchmarks.com	
Further details can be found on https://blog.cfbenchmarks.com/about/	

Disclaimer & Disclosures

- This information is provided by CF Benchmarks Ltd.
- CF Benchmarks Ltd ("CF Benchmarks") is a limited company registered in England and Wales under registered number 11654816 with its registered office at 6th Floor One London Wall, London, United Kingdom, EC2Y 5EB.
- This document describes a **draft** methodology for the CF H100 Neo-Cloud GPU Compute Index. It is provided for internal discussion purposes only, has not been approved by the Oversight Function, and does not constitute a published CF Benchmarks methodology.
- CF Benchmarks is **NOT** a registered investment advisor and does **NOT** provide investment, tax, legal, or accounting advice in any geographical locations. You should consult your own financial, tax, legal, and accounting advisors or professionals before engaging in any transaction or making an investment decision.
- All information contained within is for educational and informational purposes **ONLY**. None of the Information constitutes an offer to sell (or a solicitation of an offer to buy) any cryptoassets, security, financial product, or other investment vehicle or any trading strategy.
- Information containing any historical information, data, or analysis should not be taken as an indication or guarantee of any future performance, analysis, forecast, or prediction. Past performance does not guarantee future results.