Ornn's US H100 Compute Price Index Methodology

Ornn Research

October 11, 2025

Abstract

The Ornn US H100 Compute Price Index (HCPI) measures the real-time market cost of GPU compute using live rental offers for NVIDIA H100s across providers and regions. The index treats all offers as entries in a regional order book, weighting each price by the number of GPUs available and an exponential competitiveness factor relative to the regional median price. This up-weights abundant, competitively priced supply and down-weights overpriced or illiquid quotes. Regional indices are then combined into a global benchmark using adjusted-liquidity weights that reflect each region's effective market depth. The resulting index is path-independent, scale-invariant, and resilient to manipulation—designed to serve as a transparent, settlement-quality measure of the value of compute.

1 Introduction

The cost of high-performance compute has become a critical input to the economics of artificial intelligence. As large language models, diffusion systems, and other compute-intensive workloads scale, GPU rental markets have emerged as a key infrastructure layer linking hardware supply to AI deployment capacity. Among these, NVIDIA's H100 SXM5 GPU has rapidly become the benchmark unit of compute value. Yet despite its growing financial and operational significance, the market has lacked a standardized, high-frequency measure of H100 pricing that reflects live supply conditions across the United States.

This paper introduces a new methodology for constructing the Ornn US H100 Compute Price Index (HCPI)—a transparent, real-time measure of the fair market price of compute derived directly from executable offers. The approach departs from traditional provider-averaged frameworks by treating the market as an aggregated order book, where each quoted price level is weighted by visible GPU availability and adjusted for competitiveness relative to the regional median. This formulation captures the true distribution of marketable supply rather than nominal provider averages, ensuring that the index reflects the prices at which meaningful volume is actually offered.

By collapsing provider and listing structure into a unified price–quantity representation, the framework achieves both mathematical simplicity and economic interpretability. The exponential adjustment enforces smooth sensitivity to local price deviations, while regional aggregation by price-adjusted liquidity yields a coherent national benchmark that remains responsive to changing market depth. Together, these design choices produce an index that is stable, scale-invariant, and robust to provider concentration—positioning the Ornn HCPI as a practical settlement-grade reference for the U.S. compute market.

2 Mathematical Framework

2.1 Data Collection and Initial Processing

The system begins by collecting real-time pricing data from major cloud providers across three canonical regions: US-West, US-Central, and US-East. For each provider p offering GPU instances in region r, the raw data includes individual instance prices $\pi_{p,r}$ and corresponding GPU counts $q_{p,r}$.

2.2 Order Book Construction

Within each region, all listings are aggregated into a unified order book consisting of price–quantity pairs $(\pi_{l,r}, q_{l,r})$ where identical prices across providers are combined by summing available GPUs. By treating the market as a continuous distribution of prices across providers, this order-book formulation captures both market depth and price dispersion, offering a transparent representation of the underlying structure of compute liquidity.

2.3 Regional Index Computation

As a reference regional price, the system first computes the liquidity-weighted median price of these rental prices, m_r. Then, to account for the competitiveness of prices against this reference, the regional index assigns an exponential weight to each price level. Specifically, for a scaling hyperparameter $\lambda > 0$

$$\phi_{l,r} = \exp{(-\lambda \frac{\pi_{l,r} - m_r}{m_r})}. \tag{1}$$

The system sets $\lambda = 3$. This exponential weighting allots more importance to competitive offers, all while those prices worse than the regional median receive less weight. Then, the system computes a weighted average across the order-book, accounting for the number of GPUs offered at each price. This represents the regional index I_r.

$$I_r = \frac{\sum_l \pi_{l,r}(q_{l,r}\phi_{l,r})}{\sum_l (q_{l,r}\phi_{l,r})} \tag{2}$$

2.4 US Index Aggregation

The final US index is computed by taking a weighted average of the regional indices, where each weight is simply the price-adjusted liquidity of the order book, $G_r = \sum_l (q_{l,r}\phi_{l,r})$. In this way, the index I_{US} accounts for each region's effective market depth.

$$I_{US} = \frac{\sum_r I_r G_r}{\sum_r G_r} \tag{3}$$

3 Index Properties

3.1 Scale Invariance

The HCPI is *scale invariant*, meaning that proportional changes in all observed prices result in an identical proportional change in the index value. Formally, if each quoted price $p_{l,r}$ in the dataset is scaled by a constant factor $\alpha > 0$, then

$$I_{\mathrm{US}}(\alpha p_{l,r}) = \alpha \, I_{\mathrm{US}}(p_{l,r}).$$

This property follows directly from the definition of the exponential weighting function,

$$\phi_{l,r} = \exp\Big(- \lambda \, \frac{p_{l,r} - m_r}{m_r} \Big),$$

where both $p_{l,r}$ and the regional median m_r scale by the same factor α. Because the relative deviations $(p_{l,r} - m_r)/m_r$ remain unchanged under proportional scaling, all weights $\phi_{4l,r}$ are invariant. The index therefore scales linearly with α.

Scale invariance ensures that the HCPI measures relative price dynamics rather than nominal price levels: a uniform increase or decrease in all market quotes produces an equivalent proportional change in the index. This property preserves interpretability.

3.2 Robustness

The H100 Compute Price Index (HCPI) exhibits strong robustness to market noise and outlier quotes through two related mechanisms: *bounded influence* and *liquidity sensitivity.* These features ensure that transient or strategically posted prices exert limited impact on the benchmark while genuine, liquid price levels dominate the index computation.

Bounded Influence. Each quote's contribution to the index is exponentially attenuated according to its deviation from the regional median m_r:

$$\phi_{l,r} = \exp\Big(-\lambda \, \frac{p_{l,r} - m_r}{m_r} \Big).$$

This formulation imposes a soft upper bound on the effect of extreme or stale listings. Prices substantially above the market median receive exponentially smaller weights, while competitively priced offers are proportionally amplified. Market actors cannot manipulate the index without meaningfully changing the market itself. And unlike hard trimming or percentile filters, the exponential weighting function maintains smooth differentiability and preserves continuity as new quotes enter or leave the order book.

Liquidity Sensitivity. In addition to price-based attenuation, the weighting scheme incorporates the quantity of available GPUs $q_{l,r}$ at each price level. Larger offers—representing more tradable capacity—naturally carry greater weight, aligning the index with the effective depth of the market. This volume weighting mitigates also manipulation risk: small, unrepresentative listings cannot materially shift the benchmark without corresponding liquidity.

4 Applications and Future Development

The Ornn US H100 Compute Price Index (HCPI) provides a transparent benchmark for cash-settled derivative products and structured instruments linked to the cost of compute. Its construction from executable offers allows the index to serve as a settlement reference for futures, options, and swap contracts tied to GPU rental prices. Because the methodology decomposes the U.S. market into regional order books, it also enables geographically targeted hedging and exposure analysis—allowing traders, cloud operators, and institutional buyers to manage localized price risk across distinct compute hubs.

Future development will extend the framework along two axes. First, ongoing research will explore adaptive parameterization of the exponential weighting function using machine-learning techniques that adjust sensitivity to market conditions in real time. Additional studies will evaluate alternative weighting schemes, including entropy-based and game-theoretic approaches, to further enhance robustness and responsiveness of the benchmark under dynamic supply and demand regimes.