October 16, 2024

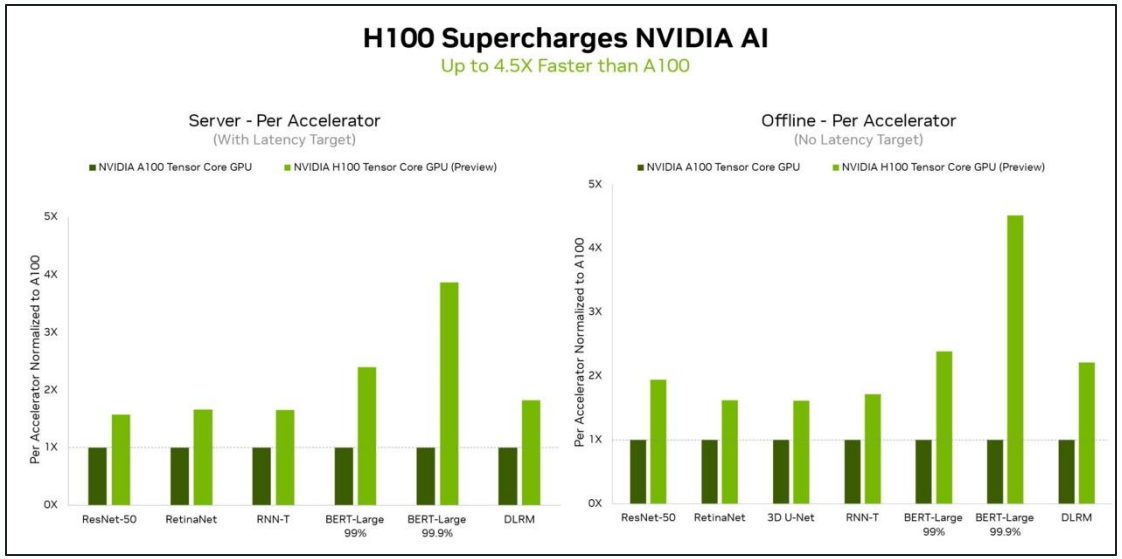
BitOoda AI Research:
The Crash of the H100 Market

Research

Tim Kelly, CEO & Founder
Dhyay Bhatt, Head of AI
Niraj Yagnik, Lead Developer & AI Engineer

Abstract
The graphics processing unit (GPU) market has undergone significant changes in recent years, driven by technological advancements in AI, shifting demand patterns, and economic factors. This report will delve into the current state of GPU pricing, focusing on the over-investment in high-end GPUs like the H100, the impact of open-source models, and the subsequent market adjustments.

Introduction
In 2023, the NVIDIA H100 GPU became the go-to choice for training large-scale AI models, with rental prices soaring to as much as $8 per hour. However, by 2024, several key developments rapidly altered this market dynamic.

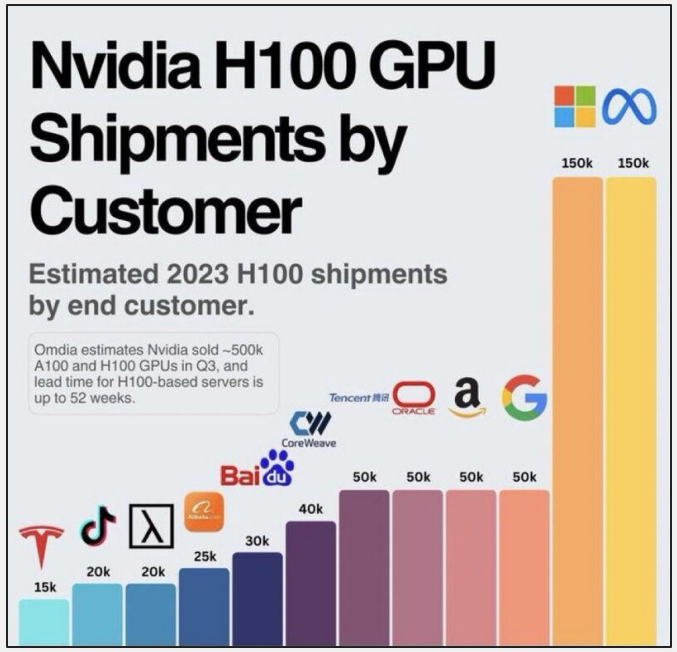


The H100 is marketed to be far superior to than NVIDIA previous generation of chips.
Source: NVIDIA

- In 2024, Meta's open-sourced 400B model led companies to shift from expensive H100 training to fine-tuning, reducing demand for H100s.
- An increased supply of H100s, coupled with the rise of cheaper alternatives for inference tasks, further pushed prices down.
- Competition from AMD and Nvidia's upcoming Blackwell GPUs will likely continue to drive H100 prices below $2 per hour.

2023
Start of the AI Hype

In 2023, the NVIDIA H100 GPU, designed specifically for **training large-scale AI models**, dominated the market with astronomical rental prices, soaring as **high as $8 per hour**. For AI startups and data centers alike, the H100 was seen as an essential tool for training cutting-edge models, and the overwhelming demand for AI training fueled both high prices and shortages in availability.



H100 GPU Shipments
Source: Omedia Research

The general sentiment was that this demand would continue to climb, and many believed the price of H100s would remain high. The H100's superior compute power, especially for AI training tasks, and the proliferation of large-scale model training created a market environment where renting H100 GPUs felt like a necessity to stay competitive in the race to develop Artificial General Intelligence (AGI). This demand was bolstered by the capital pouring into AI startups, pushing GPU clusters into a price bubble where the expectation was that H100 pricing would remain elevated for the foreseeable future.

However, that market dynamic underwent a rapid transformation in 2024, as several key events changed the landscape.

2024
The Turning Point

1.

LLAMA3.1: Meta Open Sourcing a SOTA 401B Model

A pivotal moment came when Meta open-sourced **Llama3.1,** a state-of-the-art (SOTA) 400 billion parameter model, giving companies a difficult choice: invest upwards of $100 million in building and training their own large-scale models, or fine-tune Meta's freely available model that **showcased comparable performance** to GPT's best model. **Most companies opted for the latter**, seeing the immense cost savings and shorter time to market.

This led to a sharp decline in the demand for **training resources**. With the availability of Meta's model, it became clear that many organizations no longer needed to invest in expensive GPU time for training from scratch. Instead, **fine-tuning existing models**, which requires significantly less compute power, became the more attractive and cost-effective route.

Category / Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0125)	GPT-4 Omni	Claude 3.5 Sonnet
General MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	**88.7**	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	**77.0**
IFEval	**88.6**	85.1	84.3	85.6	88.0
Code HumanEval (0-shot)	89.0	73.2	86.6	90.2	**92.0**
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	**90.5**
Math GSM8K (8-shot, CoT)	**96.8**	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	**76.6**	71.1
Reasoning ARC Challenge (0-shot)	**96.9**	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	**59.4**
Tool use BFCL	88.5	86.5	88.3	80.5	**90.2**
Nexus	**58.7**	-	50.3	56.1	45.7
Long context ZeroSCROLLS/QuALITY	**95.2**	-	**95.2**	90.5	90.5
InfiniteBench/En.MC	**83.4**	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	**100.0**	**100.0**	90.8
Multilingual Multilingual MGSM (0-shot)	**91.6**	-	85.9	90.5	**91.6**

Llama3.1 405B outperformed GPT-4o on multiple benchmark test during release
Source: @AIatMeta

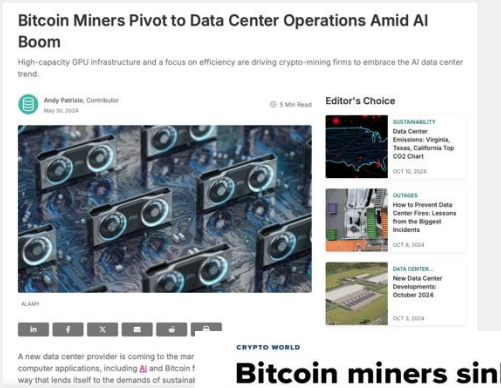
Definition of Fine-Tuning

Fine-tuning is the process of taking a pre-trained machine learning model and adapting it to a specific task by training it further on a smaller, task-specific dataset. It allows the model to leverage its existing knowledge while adjusting to new patterns in the target data, making it more effective for the specific application.

2024
The Turning Point

2.
Increased Supply of H100s

At the same time, the supply of H100s dramatically increased. Several key factors contributed to this, including better production yields and the **fulfillment of large backlogged orders from 2023**. This increased availability of GPUs made it easier for companies to acquire allocations, particularly as former mining firms pivoted into the AI space. These companies, now dubbed "**Neoclouds**," capitalized on the oversupply by entering the market with large H100 clusters just as training demand began to wane. Many of these Neoclouds focused exclusively on H100 SXMs, further adding to the supply glut.



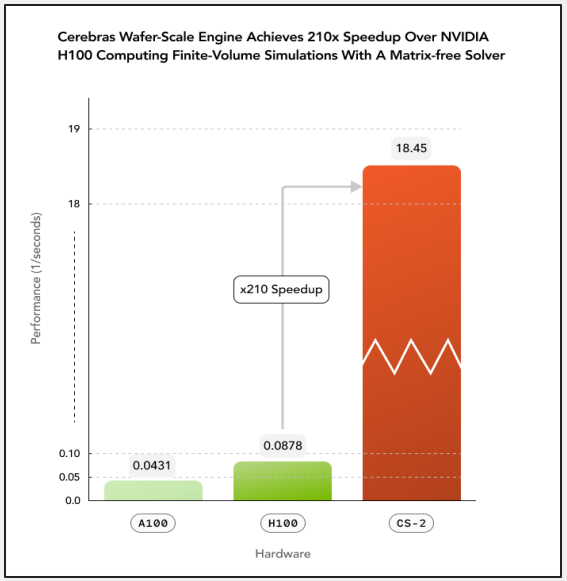
Bitcoin Miners have been pivoting to AI since early this year.

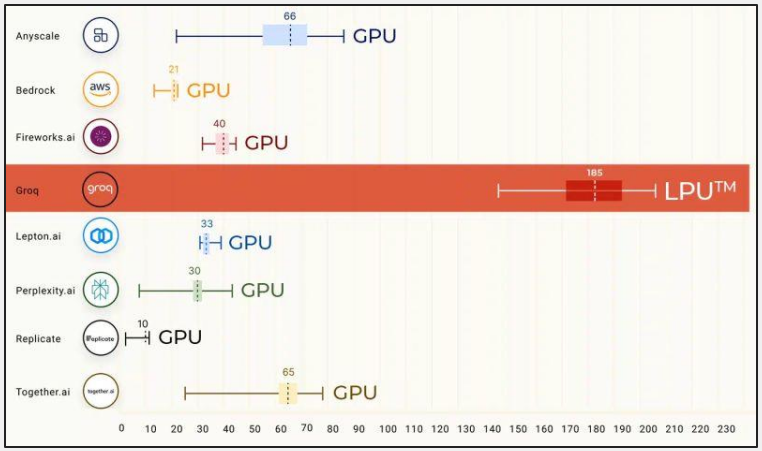
2024
The Turning Point

3.
Shift to Inference and Cheaper Alternatives

As demand for large-scale training declined, a significant portion of AI activity **shifted toward inference**, where trained models are used to make predictions and perform tasks in real time. Inference workloads require less GPU power compared to training, and cheaper alternatives to the H100—such as Nvidia's **L40 GPUs** and specialized inference hardware from companies like **Groq, Cerebras**, and **SambaNova**—became the go-to solutions.

These inference-specialized chips **offer lower cost** per compute for inference tasks, further eroding the demand for H100 GPUs. The market for GPUs was increasingly saturated, and with fewer organizations requiring the expensive H100s for training, prices began to tumble.





Groq LPU and Cerebras CS2 arecomfortably outperforming H100s

Sources: Groq and Cerebras

Future of H100 Pricing

Looking ahead

- **AMD and Blackwell Competition**: AMD's push into AI and Nvidia's upcoming Blackwell GPUs will increase competition. Blackwell is expected to outperform and underprice the H100, likely driving down H100 prices as customers upgrade.
- **Inference Hardware**: The rise of inference-specialized chips is reducing demand for H100s, as companies prefer more affordable, inference-optimized hardware.
- **Oversupply and Price Drops**: An oversupply of H100s, combined with competition and new alternatives, will likely push rental prices below $2 per hour, especially as Neoclouds expand.

Disclosures

Purpose

This research is only for the clients of BitOoda. This research is not intended to constitute an offer, solicitation, or invitation for any securities and may not be distributed into jurisdictions where it is unlawful to do so. For additional disclosures and information, please contact a BitOoda representative at info@bitooda.io.

Analyst Certification

Niraj Yagnik, the primary research analyst of this report, hereby certifies that all of the views expressed in this report accurately reflect his personal views, which have not been influenced by considerations of the firm's business or client relationships.

Conflicts of Interest

This research contains the views, opinions, and recommendations of BitOoda. This report is intended for research and educational purposes only. We are not compensated in any way based upon any specific view or recommendation.

General Disclosures

Any information ("Information") provided by BitOoda Holdings, Inc., BitOoda Digital, LLC, BitOoda Technologies, LLC or Ooda Commodities, LLC and its affiliated or related companies (collectively, "BitOoda"), either in this publication or document, in any other communication, or on or through http://www.bitooda.io/, including any information regarding proposed transactions or trading strategies, is for informational purposes only and is provided without charge. BitOoda is not and does not act as a fiduciary or adviser, or in any similar capacity, in providing the Information, and the Information may not be relied upon as investment, financial, legal, tax, regulatory, or any other type of advice. The Information is being distributed as part of BitOoda's sales and marketing efforts as an introducing broker and is incidental to its business as such. BitOoda seeks to earn execution fees when its clients execute transactions using its brokerage services. BitOoda makes no representations or warranties (express or implied) regarding, nor shall it have any responsibility or liability for the accuracy, adequacy, timeliness or completeness of, the Information, and no representation is made or is to be implied that the Information will remain unchanged. BitOoda undertakes no duty to amend, correct, update, or otherwise supplement the Information.

The Information has not been prepared or tailored to address, and may not be suitable or appropriate for the particular financial needs, circumstances or requirements of any person, and it should not be the basis for making any investment or transaction decision. The Information is not a recommendation to engage in any transaction. The digital asset industry is subject to a range of inherent risks, including but not limited to: price volatility, limited liquidity, limited and incomplete information regarding certain instruments, products, or digital assets, and a still emerging and evolving regulatory environment. The past performance of any instruments, products or digital assets addressed in the Information is not a guide to future performance, nor is it a reliable indicator of future results or performance.

All derivatives brokerage is conducted by Ooda Commodities, LLC a member of NFA and subject to NFA's regulatory oversight and examinations. However, you should be aware that NFA does not have regulatory oversight authority over underlying or spot virtual currency products or transactions or virtual currency exchanges, custodians or markets.

BitOoda Technologies, LLC is a member of FINRA.

"BitOoda", "BitOoda Difficulty", "BitOoda Hash", "BitOoda Compute", and the BitOoda logo are trademarks of BitOoda Holdings, Inc.