BitOoda

BitOoda AI Research:
An Educational Analysis of Inference

Abstract
In the realm of artificial intelligence (AI), understanding key performance metrics such as throughput and latency is essential for selecting appropriate hardware for AI inference tasks. This educational research report introduces these fundamental concepts, explains the principles of AI inference, and analyzes current hardware offerings, including specialized processors and NVIDIA's H100 GPUs. By examining performance data for various processors on the Llama 3.1 models (70 billion (70B) and 8 billion (8B) parameters), we highlight how specialized hardware can offer superior performance in terms of throughput and latency compared to general-purpose GPUs for inference.

Introduction
The exponential growth of AI applications has led to increasing demand for computational resources. Performance metrics like throughput and latency have become critical factors in the deployment of AI models, especially during the inference phase. Inference, the process of using a trained model to make predictions on new data, requires efficient hardware to meet real-time processing needs.

This report aims to educate readers new to AI on the importance of throughput and latency in AI inference. We delve into the first principles of AI inference, compare specialized hardware solutions with general-purpose GPUs like the NVIDIA H100, and analyze performance data to demonstrate how these concepts apply in practice.

Research

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

Key Takeaways

- Specialized hardware can offer advantages over general-purpose GPUs (e.g., NVIDIA H100) in AI inference tasks, particularly for larger models.
- Benefits may include improved throughput and reduced latency, contributing to more efficient and responsive AI applications.
- Understanding throughput and latency is important for making informed hardware choices tailored to specific use cases.
- As AI models increase in size and complexity, the role of optimized hardware may become more significant.

Definitions and Background
More on Inference

Understanding AI Inference

What is AI Inference?
AI inference applies a trained model to new data to generate predictions or outputs. Unlike training, which adjusts model parameters, inference uses fixed parameters to compute results on input data.

Inference Process
- *Input Processing:* Raw data is preprocessed, e.g., text tokenized for language models.
- *Model Computation:* The input passes through the model, using learned weights and biases to generate outputs.
- *Output Generation:* The output is post-processed into a readable format, like converting token probabilities to words.

Factors Affecting Performance
- ***Model Size:*** Larger models (e.g., 70B vs. 8B parameters) require more resources, affecting latency and throughput.
- ***Hardware Efficiency***: Specialized hardware enhances performance compared to general-purpose GPUs.
- ***Software Optimization:*** Efficient frameworks improve memory and computational efficiency.

Important Definitions

Throughput
Throughput refers to the amount of data processed in a given time frame. In AI language models, it is often measured in ***tokens per second***. High throughput is crucial for applications that require processing large amounts of data quickly, such as batch processing or generating long text sequences.

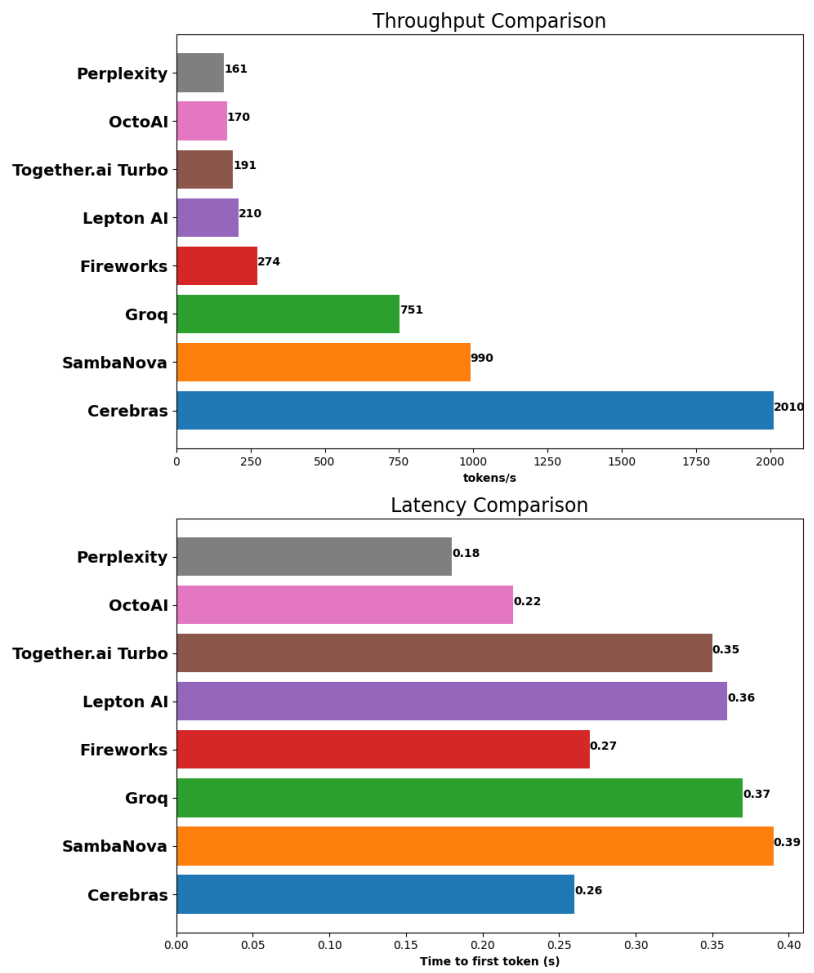
Example: A model generating 1,000 tokens per second has higher throughput than one generating 500 tokens per second.

Latency
Latency is the time delay from the initiation of a request to the delivery of the response. In AI inference, it is commonly measured as ***Time to First Token (TTFT)***, indicating how quickly the model begins to generate output after receiving an input.
- **Lower latency** is essential for real-time applications like interactive chatbots or live translations.
- **Higher latency** may be acceptable in scenarios where immediate responses are not critical.

Hardware Analysis
Specialized Hardware vs. NVIDIA H100 GPUs

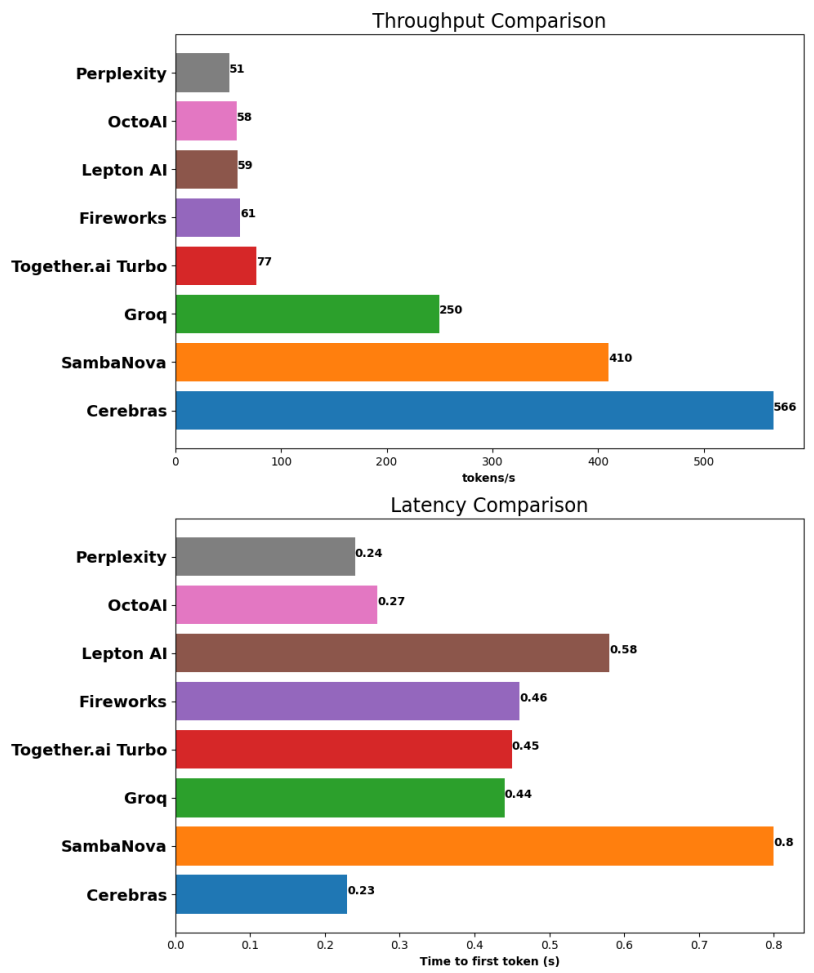
Advancements in AI hardware have led to the development of specialized processors designed specifically for AI workloads. These include hardware from companies like Cerebras, SambaNova, and Groq. NVIDIA's H100 GPUs, while powerful, are general-purpose and may not offer the same level of optimization for specific AI tasks. Cerebras delivers the highest throughput and lowest latency for both models

LLAMA 3.1 8B



Source: Artificial Analysis (https://artificialanalysis.ai/)

LLAMA 3.1 70B



Source: Artificial Analysis (https://artificialanalysis.ai/)

Specialized Hardware Advantages and Implications

Is specialized hardware always better for inference?

Advantages

Architectural Optimization
Specialized hardware is architected specifically for AI workloads. This means that the hardware components are optimized for the matrix and tensor operations commonly used in neural networks, leading to improved performance.

Scalability
These hardware solutions are designed to scale efficiently with model size. They handle larger models without a proportional increase in latency, which is essential as AI models continue to grow in complexity.

Energy Efficiency
Specialized processors often consume less power per computation compared to general-purpose GPUs. This energy efficiency can lead to cost savings in large-scale deployments.

Implications for AI Practitioners

Application-Specific Hardware Selection
Understanding the trade-offs between throughput and latency is crucial when selecting hardware for AI applications.

- ***High Throughput Needs***: Applications that require processing large volumes of data quickly, such as batch data processing or generating extensive text outputs, would benefit from hardware with high throughput.
- ***Low Latency Requirements***: Real-time applications like conversational AI, live translations, or interactive systems need hardware that minimizes latency to provide immediate responses.

Cost Considerations
While specialized hardware may have higher initial costs, the performance benefits can lead to lower total cost of ownership due to reduced energy consumption and the need for fewer units to achieve the desired performance.

Software Ecosystem
Selecting hardware supported by robust software frameworks is essential. Compatibility with popular AI libraries and tools ensures easier integration and development.

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.