Ian Loe

THE FUTURE OF A.I. PROCESSORS

COMPARING THE HUAWEI ASCEND 910C VS THE NVIDIA H100

Summary

The comparison between the Huawei Ascend 910C and the NVIDIA H100 represents a pivotal moment in the A.I. semiconductor landscape, particularly in the context of global trade tensions and technological rivalry. The Ascend 910C, developed by Huawei, is positioned as a formidable alternative to NVIDIA's H100, which is widely recognised for its high performance and efficiency in A.I. and high-performance computing applications. The emergence of the Ascend 910C is driven by U.S. trade restrictions limiting access to advanced chips in China, presenting Huawei with a unique opportunity to capture a significant share of the A.I. market. The Ascend 910C aims to deliver performance comparable to the H100, with expectations of strong demand highlighted by projected orders exceeding 70,000 units. Despite challenges during its development, including manufacturing hurdles that led to reduced specifications from its predecessor, the Ascend 910B, the chip signifies Huawei's commitment to advancing semiconductor technology amidst increasing regulatory pressures. In contrast, the NVIDIA H100 leverages advanced architectural innovations, such as fourth-generation Tensor Cores and enhanced memory capabilities, to achieve substantial performance gains, particularly in A.I. inference and data analytics. Notably, both chips are designed to support a range of applications, from data centres to edge computing solutions. However, the competitive landscape is marked by ongoing debates about total cost of ownership (TCO), ecosystem maturity, and market accessibility. While the H100 has been praised for its performance efficiency and extensive software support, concerns regarding Huawei's security practices and corporate governance continue to cloud the Ascend 910C's reception in certain markets. This dynamic interplay of technological capability and ethical considerations shapes the future of A.I. hardware development and deployment across the globe.

Huawei Ascend 910C

Performance Comparisons

Huawei has asserted that the capabilities of the Ascend 910C are comparable to those of NVIDIA's H100, a chip that has gained substantial global attention for its high performance. The Ascend 910C aims to fill the gap left by the H100's unavailability, marking a potential breakthrough for Huawei amid challenges posed by U.S. sanctions on semiconductor. The expected order volume for the 910C is projected to exceed 70,000 units, indicating strong market interest and confidence in its performance.

Technical Specifications

Originally intended to double the performance of its predecessor, the Ascend 910B, the Ascend 910C was designed with an increase in the number of active A.I. cores. However, due to manufacturing challenges, Huawei ultimately reduced the specifications to align more closely with the existing capabilities of the 910B. Despite this, the 910C still represents a significant step forward, building on the theoretical maximum performance upgrades established in previous iterations. The first-generation Ascend 910 boasted a performance of 320 TFLOPS, while the second-generation 910B achieved 400 TFLOPS, setting the stage for expectations surrounding the 910C.

Market Context and Strategic Importance

As Huawei navigates intense competition and regulatory hurdles, the development of the Ascend 910C underscores the company's commitment to advancing its semiconductor capabilities. The chip is being tested with various prominent Chinese internet and telecom firms, including ByteDance and Baidu, to assess its performance and market viability. This strategic positioning is crucial as the global semiconductor landscape continues to evolve, with Huawei striving to carve out a significant share amidst external pressures and internal challenges.

Nvidia H100

The NVIDIA H100 GPU, part of the NVIDIA Hopper architecture, is designed for high-performance computing (HPC), artificial intelligence (A.I.), and data analytics applications. It significantly accelerates over 4,000 applications and is suited for various environments, ranging from data centres to edge computing solutions. The H100 offers dramatic performance enhancements and cost-saving benefits, making it a versatile tool for a wide array of workloads.

Architectural Innovations

The H100 incorporates advanced features that enhance its performance in A.I. and HPC tasks. It utilises fourth-generation Tensor Cores, which support a variety of data types including FP64, TF32, FP32, FP16, INT8, and FP8. These enhancements optimise memory usage and deliver up to 30 times higher inference speeds while minimising latency, crucial for demanding A.I. applications. Additionally, the introduction of the Tensor Memory Accelerator (TMA) and asynchronous execution capabilities enables better overlap of data movement, computation, and synchronisation, improving overall GPU utilisation.

Performance Metrics

In benchmarks, the H100 has shown substantial improvements over its predecessor, the A100. It is reported to deliver double the raw dense and sparse matrix math throughput per Streaming Multiprocessor (SM), and its new DPX instructions allow for up to 7 times better performance in dynamic programming algorithms, essential for fields such as genomics and robotics. In financial computing scenarios, the H100 has achieved groundbreaking performance metrics, such as sub-10ms warm times in benchmark tests, which demonstrates its efficiency in real-time applications.

Power and Availability

The H100 is available in two main form factors: the SXM mezzanine for high-performance servers and a PCIe card for mainstream server applications. The SXM variant has a thermal design power (TDP) of 700 watts, a significant increase compared to the A100's 400 watts, reflecting its enhanced capabilities. The previously high-demand H100 data center GPU has seen a reduction in delivery wait times from a peak of 8-11 months to 3-4 months, indicating a

relief in supply pressure. Additionally, with major cloud providers such as AWS, Google Cloud, and Microsoft Azure offering easier access to A.I. computing services for customers, enterprises that previously purchased large quantities of H100 GPUs have begun further reselling these GPUs.

Software Integration

To facilitate the adoption of the H100, NVIDIA offers a five-year subscription to the NVIDIA A.I. Enterprise software suite, which includes enterprise support. This comprehensive suite simplifies the deployment of A.I. workflows across various sectors, such as chatbots, recommendation systems, and vision A.I.

Comparative Analysis

Performance Metrics

The performance of A.I. hardware can be quantitatively assessed through benchmarks such as MLPerf Inference 3.0, which evaluates A.I. performance across various real-world applications. The Nvidia H100, equipped with the Hopper architecture and Transformer Engine, has demonstrated remarkable performance in these benchmarks, achieving significant efficiency gains in A.I. inference tasks, including generative .AI. applications. For instance, the H100 GPUs showcased a performance increase of up to 54% compared to previous generations, highlighting its capability in high-demand scenarios where quick responses are essential. In contrast, the Huawei Ascend 910C has also been optimised for A.I. workloads but may not consistently match the H100's performance metrics in similar tests. While it boasts substantial computing capabilities, its relative performance against leading competitors like the H100 can depend heavily on the specific workloads and the optimisation level of the software stack utilised.

Total Cost of Ownership (TCO)

TCO is a critical factor for organisations considering A.I. infrastructure. The Nvidia H100 is designed to provide substantial reductions in TCO through improved performance and energy efficiency. For example, state-of-the-art large language models (LLMs) can achieve up to a 5.6x reduction in energy consumption compared to prior models, thanks to optimisations

inherent in the H100 architecture. This reduction in operational costs can be particularly appealing for enterprises scaling their A.I. capabilities. While Huawei's solutions, including the Ascend 910C, aim for competitive TCO through energy-efficient designs, specific comparisons on cost-effectiveness may vary based on deployment size and operational strategies. The lack of comprehensive benchmarking against Nvidia's offerings makes it challenging to definitively state how the Ascend 910C stacks up in terms of TCO.

Ecosystem and Software Support

Nvidia has developed a robust ecosystem for its hardware, with tools like TensorRT and Triton Inference Server designed to enhance performance and streamline development for A.I. applications. These resources facilitate not only ease of use but also effective optimisation strategies that can significantly boost the performance of A.I. models running on Nvidia hardware. Conversely, Huawei's ecosystem for the Ascend 910C includes its own frameworks and tools tailored for A.I. development. However, the depth of software support and community engagement seen with Nvidia's CUDA and related technologies may pose a challenge for Huawei in terms of widespread adoption and developer familiarity. The differences in ecosystem maturity can influence the decision-making process for organisations evaluating which hardware to implement.

Application in A.I. and Machine Learning

The advancements in A.I. and machine learning are reshaping various industries, with significant contributions from both the Huawei Ascend 910C and NVIDIA H100 chips. These processors are engineered to enhance the efficiency and capabilities of A.I. applications across different sectors, including finance, healthcare, and manufacturing.

A.I. Use Cases

The application of A.I. has become a priority across multiple domains, particularly in automating processes and improving decision-making. For instance, the U.S. Postal Service utilises NVIDIA's Edge Computing Infrastructure Program to support numerous A.I. applications powered by the EGX™-based systems, demonstrating the chip's versatility in real-world scenarios. Moreover, A.I. high performers, those organisations realising at least 20% of their earnings before interest and taxes (EBIT) from A.I., leverage generative A.I. across diverse functions, including product development and risk management. Huawei envisions A.I. as a general-purpose technology, akin to the transformative impacts of railroads and electricity in

the 19th century. The Ascend 910C is designed to facilitate broad deployment across cloud, edge computing, and IoT devices, thereby fostering a collaborative ecosystem for A.I. development. Its application capabilities extend beyond traditional machine learning to encompass advanced solutions, enhancing efficiency and driving innovation in various sectors.

Comparative Performance

Both the Ascend 910C and NVIDIA H100 are positioned to serve high-demand A.I. environments. Reports suggest that the Ascend 910C's performance is comparable to that of the H100, particularly in machine learning tasks that require substantial computational power. This positioning could potentially disrupt NVIDIA's market leadership by filling the gap created by U.S. trade restrictions on advanced semiconductor technologies in China.

Efficiency and Resource Utilisation

In terms of algorithm development, both companies emphasise the importance of efficiency. Huawei's strategy focuses on creating data-efficient and energy-efficient algorithms that can deliver superior results with minimised resources. Similarly, NVIDIA's A.I. software stack accelerates production timelines for A.I. projects, highlighting the need for effective resource utilisation in the development process.

Cost Analysis

The cost dynamics between the Huawei Ascend 910C and Nvidia H100 GPUs reflect a complex interplay of production expenses, market demand, and regulatory influences.

Pricing Overview

As of early 2024, the pricing for Nvidia's H100 GPUs varies significantly depending on the market. Analysts estimated that these GPUs are selling for approximately $25,000 to $30,000 each in the U.S., with some individual units being sold on platforms like eBay for over $40,000. Conversely, the pricing details for the Huawei Ascend 910C remain less transparent but are suggested to be competitive, particularly as interest from major firms like ByteDance has been indicated to surpass production capacities, suggesting robust demand.

Total Cost of Ownership

When analysing costs, customers adopting A.I. technologies often consider the total cost of ownership (TCO), which encompasses more than just the initial price of the GPU. Factors

include the cost of additional necessary infrastructure, such as GPU servers, management head nodes, networking equipment, storage solutions, data centre IT staff, maintenance, and operational expenses related to power consumption and space rental. Notably, high-performance GPUs like the H100 and Ascend 910C can lead to significant performance speedups, ultimately reducing these operational costs over time, which should be factored into any comprehensive cost analysis.

Regulatory and Market Influences

The competitive landscape has also been shaped by recent U.S. export restrictions that specifically target high-performance computing components intended for the Chinese market. For instance, chips like the H100 are subjected to stringent performance thresholds that affect their availability in certain regions. This regulatory environment could further influence pricing and demand, particularly for Huawei's offerings, which may provide a more accessible alternative in markets restricted from accessing Nvidia’s products.

Ethical Considerations and Controversies

The comparison between the Huawei Ascend 910C and the Nvidia H100 is deeply intertwined with ethical considerations, particularly concerning security, diversity, and corporate governance.

Security Concerns

A significant issue surrounding Huawei, and by extension the Ascend 910C, is the security of its products. Numerous governments, particularly in the United States and Australia, have raised alarms about potential espionage capabilities associated with Huawei technology. U.S. intelligence agencies have warned that the Chinese government could leverage Huawei to spy on foreign nations, citing vague Chinese laws that may compel companies to cooperate with state intelligence operations. The Ascend 910C's integration in critical infrastructure raises concerns about backdoors that could allow unauthorised data access, potentially putting sensitive information at risk. Furthermore, incidents such as a software update containing malicious code that targeted Australia’s telecommunications network highlight the potential vulnerabilities linked to Huawei devices.

Comparisons with Nvidia

Nvidia, while not without its controversies, particularly related to antitrust allegations, operates within a different ethical framework. Its products are scrutinised not just for performance but also for their implications in A.I. and data security. As Nvidia continues to solidify its position as a leader in A.I. technologies, concerns arise about monopolistic practices and their broader impact on competition in the tech sector. The ethical considerations surrounding Nvidia's business practices often revolve around maintaining fair competition while delivering cutting-edge technologies.