Research Brief:

INSIDE THE QUALCOMM CLOUD AI 100-SERIES INFERENCE ACCELERATORS

STEVE MCDOWELL, CHIEF ANALYST

JANUARY 2024

INSIDE THE QUALCOMM CLOUD AI 100-SERIES INFERENCE ACCELERATORS

CONTEXT

Qualcomm has made significant strides in AI inference processing with its Cloud AI 100 series accelerators, which Qualcomm recently expanded with the announcement of its Qualcomm Cloud AI 100 Ultra.

Qualcomm's cutting-edge accelerators are engineered to enhance AI capabilities in cloud and edge computing, marking Qualcomm's ambitious entry into the high-demand AI market.

This Research Brief unpacks the features, applications, and technological advancements of the Qualcomm Cloud AI 100 and Cloud AI 100 Ultra.

QUALCOMM CLOUD AI 100

The Qualcomm Cloud AI 100, introduced in 2020, is an Artificial Intelligence (AI) inference accelerator. It's designed to enhance cloud computing environments by offering efficient, high-speed AI inference processing capabilities.

The key aspects of the Qualcomm Cloud AI 100 include:

- **Purpose-Built for AI Inference**: The Cloud AI 100 is specifically engineered for AI inference tasks, which involve applying trained AI models to new data to make predictions or decisions.
- **High Performance**: The AI 100 is built to deliver high throughput and efficiency for AI inference workloads, essential for applications such as speech recognition, natural language processing, image and video analysis, and other AI-driven services.
- **Energy Efficiency**: One of the key features of the Qualcomm Cloud AI 100 is its energy efficiency. The device is designed to provide high AI processing power while consuming less energy, crucial for reducing operational costs and environmental impact in data centers.
- **Scalability**: The architecture of the AI 100 allows for scalability, enabling it to cater to various demands of cloud services, from small-scale applications to large, complex AI models and workloads.
- **Broad Applicability**: It's suited for various industries and applications, including healthcare, automotive, retail, and smart cities, where AI inference is rapidly growing in importance.
- **Software Support**: Qualcomm provides software support for the Cloud AI 100, including tools and frameworks that facilitate the deployment and optimization of AI models on this platform.

© Copyright 2024 NAND Research

VARIANTS

The Qualcomm Cloud AI 100 caters to various deployment scenarios with a range of form factors tailored to various server configurations and use cases, particularly in cloud and edge computing environments.

While specific details about each form factor may vary, they generally include the following types:

Form Factor	Description	Ideal Use Case
PCIe Card	Standard PCIe card format for easy integration into server architectures.	Data center servers, general AI tasks.
PCIe Lite Accelerator	Configurable 35-55W TDP, optimized for AI Edge market, balancing power and performance.	Energy-efficient edge computing.
M.2 Form Factor	Compact size ideal for embedded systems and IoT devices where space is a constraint.	IoT devices, embedded AI applications.
Datacenter Platforms	Custom server configurations tailored for traditional data center setups.	Large-scale cloud deployments.
Edge Computing Platforms	Configurations specifically designed for edge computing environments.	Edge devices, real-time AI processing.

These form factors enable the Qualcomm Cloud AI 100 to be versatile and adaptable for a wide range of applications, from large data centers to space-constrained edge environments.

Form Factor	HHHL-Pro	HHHL-Std	HHHL-Lite	Dual M.2	Dual M.2e
Power (TDP)	75 W	75 W	35, 45, 55 W (Configurable)	15 W – 25 W	15 W – 25 W
ML Capacity (INT8)	Up to 400 TOPS	Up to 350 TOPS	Up to 137 TOPS	Up to 200 TOPS	Up to 70 TOPS
On-Die SRAM	144 MB	126 MB	81 MB	126 MB	81 MB
On-Card DRAM	32 GB LPR4x 137 GB/s	16 GB LPR4x 137 GB/s	16 GB LPR4x 137 GB/s	16 GB LPR4x 137 GB/s	8 GB LPR4x 68 GB/s
Host Interface	PCIe Gen 4 8 lanes	PCIe Gen 4 8 lanes	PCIe Gen 4 8 lanes	PCIe Gen 3 4 lanes	PCIe Gen 3 4 lanes

© Copyright 2024 NAND Research

These configurations are designed to make the Cloud AI 100 adaptable to a wide range of applications, from large-scale cloud deployments to power-sensitive edge computing scenarios. This adaptability ensures that the Cloud AI 100 can meet the diverse needs of AI inference processing across different industries and use cases.

MLPERF 3.1 RESULTS

The MLCommons MLPerf benchmarks are the most widely accepted benchmarks for measuring the performance of machine learning (ML) hardware, software, and systems. The primary goal of MLPerf is to provide fair, reliable benchmarks for evaluating the performance of different ML and AI systems and components across various deployment scenarios and applications.

In September 2023, MLCommons released[1] its MLPerf Inference 3.1 benchmark results, in which Qualcomm demonstrated significant advancements with its Cloud AI 100 inference accelerators.

The results show notable improvements in performance, power efficiency, and lower latencies, particularly for Natural Language Processing (NLP) and computer-vision networks.

- **Performance and Power Efficiency Improvements**: Qualcomm's submissions for the MLPerf Inference v3.1 benchmarks surpassed its previous records. In several categories, the Cloud AI 100 showed advancements in peak offline performance, power efficiency, and latency reduction. For instance, a 2U datacenter server platform equipped with 16 Qualcomm Cloud AI 100 PCIe Pro (75W TDP) accelerators displayed a 15-20% improvement in power efficiency across NLP and computer vision networks.

- **RetinaNet Network Optimization**: The performance of the RetinaNet Network on platforms utilizing the Cloud AI 100 has been optimized by approximately 12%. This optimization indicates Qualcomm's continued efforts to enhance AI models' processing efficiency and speed.

- **New Collaborations and Server Platforms**: Qualcomm expanded its collaborations, introducing new server platforms for edge and data center categories. Partners such as Lenovo and HPE have integrated the Cloud AI 100 into their server platforms, like the ThinkSystem SR665v1 and HPE ProLiant servers, showcasing the adaptability and broad applicability of the Cloud AI 100.

- **Extension to Network Division**: Qualcomm also extended its MLPerf submissions to include the RetinaNet Network in the Network division, achieving results nearly equivalent to those in the Closed division. This extension demonstrates the Cloud AI 100's versatility in different deployment conditions.

The MLPerf Inference v3.1 results clearly demonstrate the effectiveness of the Qualcomm Cloud AI 100 across a broad range of applications, including both edge and data center categories, highlighting its performance in key metrics like inference-per-second and inference-per-second-per-watt (I/S/W).

[1] MLCommons, September 9, 2023: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/

© Copyright 2024 NAND Research

NEW: QUALCOMM CLOUD AI 100 ULTRA

In November 2023, Qualcomm added to its Cloud AI 100 lineup with the introduction of its new

Qualcomm Cloud AI 100 Ultra. The new accelerator is tailored explicitly for demanding AI tasks, particularly on generative AI and large language models.

The new accelerator offers four times the performance of its earlier Cloud AI 100 variants. It supports models up to 100 billion parameters on a single 150-watt card and 175 billion parameters with two cards.

Form factor:	**PCIe FH3/4L**
TDP:	**150W**
ML capacity (INT8):	**870 TOPs**
On-die SRAM:	**576 MB**
On-card DRAM:	**128 GB LPR4x 548 GB/s**
Host interface:	**PCIe Gen 4, 16 Lanes**
Number of cores:	**64 AI cores on single card**

The key features and capabilities of the Qualcomm Cloud AI 100 Ultra can be summarized as follows:

- **Enhanced Performance**: The Qualcomm Cloud AI 100 Ultra delivers substantially higher performance than previous generations.
- **Large Model Support**: The AI 100 Ultra can support extremely large AI models, handling models with up to 100 billion parameters on a single 150-watt card. The Ultra can scale up to support 175 billion parameter models with two cards. Multiple AI 100 Ultra cards can be combined to handle even larger models.
- **Programmability and Flexibility**: A key aspect of the AI 100 Ultra is its programmability, allowing it to adapt to the latest AI techniques and data formats. This flexibility ensures that it can keep pace with the rapidly evolving field of artificial intelligence.
- **Integration with Qualcomm AI Stack and Cloud AI SDK**: The device is designed to work seamlessly with Qualcomm's AI Stack and Cloud AI SDK, facilitating the development and deployment of AI models and enhancing the overall AI ecosystem.
- **Energy Efficiency and Sustainability**: Despite its high performance, the AI 100 Ultra maintains energy efficiency. This characteristic is crucial for reducing operational costs in data centers and supports sustainability goals in AI operations.

© Copyright 2024 NAND Research

- **Cost-Effectiveness**: The AI 100 Ultra offers an impressive performance per total cost of ownership (TCO) dollar, making it an attractive option for businesses looking to leverage advanced AI capabilities without incurring prohibitive costs.

- **Broad Applicability**: Like its predecessor, the AI 100 Ultra is suitable for various applications across various industries, including healthcare, automotive, retail, and others where advanced AI processing is needed.

Its blend of price-performance, power efficiency, scalability, and security positions the Qualcomm Cloud AI 100 Ultra as an excellent choice for organizations seeking to leverage advanced AI while also focusing on sustainability.

ADOPTION

Qualcomm's Cloud AI 100 solutions offer high-performance, low-power deep learning inference acceleration. Its scalable architecture enables several real-world AI applications:

- **Predictive Applications**: The Qualcomm Cloud AI 100 solutions support a range of predictive applications, particularly in deep learning. The accelerators can analyze large datasets to provide valuable predictions, identify trends, and enhance cybersecurity by protecting against cyberattacks.

- **Real-time Scene Analysis**: The Cloud AI 100 Edge Development Kit, with its low power consumption and high performance, can be used for continuous monitoring and analysis. In retail environments, it can detect high-risk areas for theft, identify suspicious activities or hazards like spills, and analyze shopper behavior to offer recommendations for enhancing the shopping experience.

- **Web Feed Analysis**: Cloud AI 100 systems optimize user experiences on the web. They enable real-time translation of web posts, filter explicit content using natural language processing, and provide more relevant content suggestions to users.

- **Text to Code Applications**: The Cloud AI 100 powers applications like CodeGen, which converts natural language conversations into high-quality programming code. This showcases the capability of Qualcomm's AI solutions in AI inference acceleration, facilitating efficient and innovative coding processes.

Qualcomm's AI solutions are versatile, enabling advanced applications in various fields, from healthcare to retail, web optimization, and software development.

PLATFORM PARTNERS

The Qualcomm Cloud AI 100 accelerator family is available from several tier-one technology providers including Lenovo, Hewlett Packard Enterprise (HPE), Inventec, Foxconn, Gigabyte, and Asus.

© Copyright 2024 NAND Research

CLOUD AVAILABILITY

Amazon Web Services (AWS) recently introduced its first Qualcomm-based accelerated instance type, the DL2q, which uses the Qualcomm Cloud AI 100. While the new instance type can be used for general inference applications, the companies have highlighted its applicability is developing automotive ADAS and related applications – an area in which Qualcomm is rapidly expanding its presence.

The Cloud AI 100 accelerator is also available from Cirrascale Cloud Services as part of its CIrrascale AI Innovation Cloud.

ANALYSIS

AI inference is becoming a critical functionality, especially with large language models. The Qualcomm AI 100 and its enhanced version, the Cloud AI 100 Ultra, represent a significant leap in AI inference technology, particularly catering to the burgeoning needs of cloud and edge computing.

The Cloud AI 100, with its robust processing capabilities and energy efficiency, is adept at handling diverse and large-scale AI models, a critical factor in the rapidly expanding field of AI applications. Its focus on low latency and high throughput makes it ideal for real-time processing tasks.

Qualcomm's accelerators take these capabilities further, specifically targeting the demands of generative AI and large language models. This advanced version stands out for its ability to support extremely large AI models, demonstrating Qualcomm's commitment to pushing the boundaries of AI technology.

By delivering enhanced performance and maintaining energy efficiency, the Cloud AI 100 Ultra offers a compelling solution for complex AI tasks while keeping operational costs in check.

These products underscore Qualcomm's strategic move into high-end AI inference markets, showcasing their potential to reshape AI processing in various industries, from healthcare to automotive and beyond.

© Copyright 2024 NAND Research

© Copyright 2024 NAND Research.

NAND Research is a registered trademark of NAND Research LLC, All Rights Reserved.

This document may not be reproduced, distributed, or modified, in physical or electronic form, without the express written consent of NAND Research. Questions about licensing or use of this document should be directed to info@nandresearch.com.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nandresearch.com or visit our website at nandresearch.com.