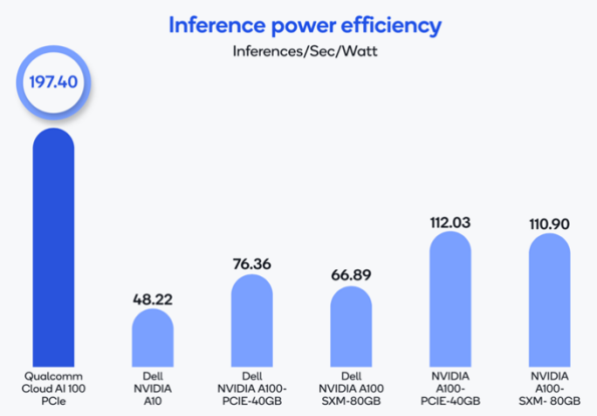
Qualcomm Technologies Delivers Top Performance and Power Efficiency in AI Benchmarks

COMPANY'S MOBILE HERITAGE MEANS EVERY MILLIWATT COUNTS.

Qualcomm, perhaps best known for its leadership Snapdragon mobile platform, has further enhanced its AI bona fides with V1.1 MLPerf inference benchmark suite released in September. Not only did the company run a clean sweep against all other platforms in terms of power efficiency (e.g., images/second/watt), the Cloud AI 100 won the best overall performance rank for object detection. Specifically, Qualcomm was able to deliver the highest data center inference performance for image processing, the highest performance per watt across the board and the fastest edge performance. Realizing up to 34% improvement in performance since the last MLPerf V1.0 release, Qualcomm has now vaulted from a niche player to a broad provider of AI acceleration platforms. Qualcomm submitted 82 benchmarks results including 36 power results and expanded coverage including language processing (BERT). Let's take a closer look.

THE QUALCOMM CLOUD AI 100 IN THE DATA CENTER

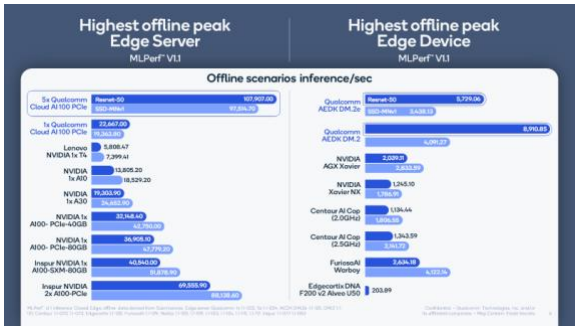
Qualcomm submitted results for online and offline runs of image processing with ResNet50, SSD Large and Small as well as BERT Large. Qualcomm delivered the highest performance efficiency of all entrants in all models and platforms. In addition, a Gigabyte server with 16 Cloud AI 100 cards delivered the highest inference performance of all submissions, beating an Inspur and NVIDIA DGX server with 8 x A100 GPUs each.



The Cloud AI 100 consumes only 15-75 watts, compared to 300-500 watts of power consumed by each GPU. So, on a chip-to-chip basis, the Qualcomm AI 100 delivers 50% of the performance at only 15% of the power. Compared to a Dell server[1] with Nvidia A100 GPU, the Qualcomm advantage rises to 2.6 X better performance per watt at a complete system level, not just at a chip level.[2]

LEADERSHIP EDGE PERFORMANCE, LATENCY AND EFFICIENCY

While power efficiency in the data center is important, it is vitally so on the edge and edge server configurations, and here the Qualcomm technology has no peer. Looking at the AI Edge Development Kit (AEDK) development platform (using a DM.2e 15-watt version), computing applications realize a full four times better power efficiency for edge devices, and twice the efficiency in edge cloud server applications tested. [3]

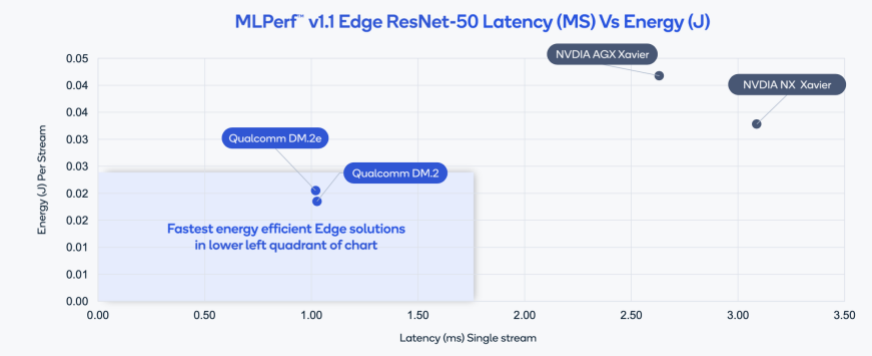


[1] Dell 1.1-006 Vs Qualcomm 1.1-058 (197.4/76.36 = 2.585).

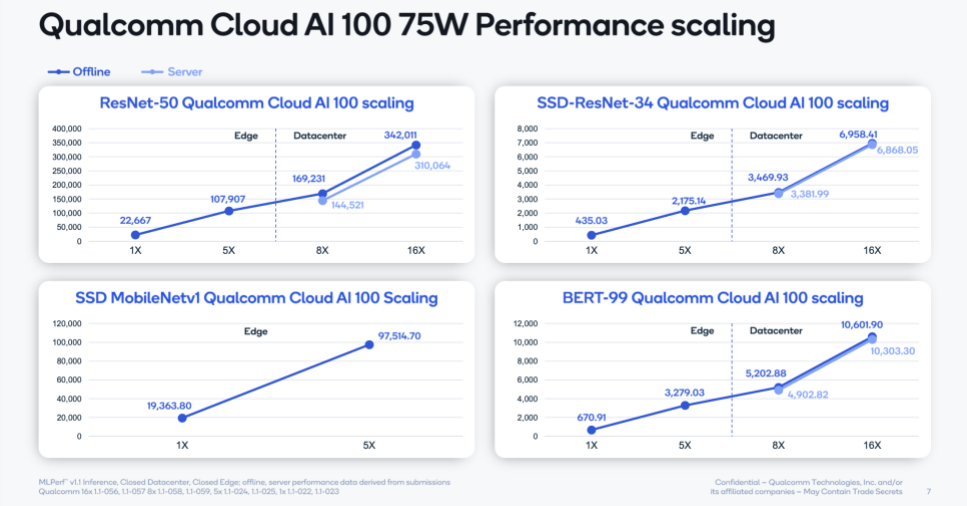
[2] MLPerf ™v1.1 Inference, Closed Datacenter Power; Power efficiency data derived from power submissions and it's not a MLPerf™ primary metric Submission Ids: Qualcomm 1.1-058, Dell: 1.1-006, 1.1-014, 1.1-016, Nvidia 1.1-037, 1.1-051

[3] MLPerf™ v1.1 Inference Closed Edge; offline data derived from Submissions: Edge server Qualcomm 1x 1.1-022, 5x 1.1-024; AEDK DM.2e 1.1-120, DM.2 1.1-121; Centaur 1.1-072, 1.1-073, Edgecortix 1.1-128, FuriosaAI 1.1-129, Nvidia 1.1-100, 1.1-109, 1.1-103, 1.1-104, 1.1-110, 1.1-117, Inspur 1.1-077 1.1-083

Qualcomm is also delivering the lowest latency[4] at the lowest power, critical for many computer vision applications such as autonomous guidance, on premise safety and security applications that demand the lowest response time to be able to react to rapidly changing scenes.



Finally, we note that Qualcomm demonstrated excellent near-linear scaling for image and language processing. A platform that can scale well, and deliver much better efficiency without sacrificing throughput, will be very attractive to data centers looking to deliver cost-effective inference processing.[5]



CONCLUSIONS

With industry leading advancements in performance density and performance per watt, the Qualcomm Cloud AI 100 platforms are leading in all the latest benchmarks. This is a major shift from a world dominated by a single vendor, and the addition of power metrics, measured at the wall, is a great step forward for clients wanting to understand Total Cost of Ownership and those who have taken a Carbon Pledge. Qualcomm has significantly expanded its submission to MLPerf benchmarks, doubling the number of platform submissions from edge to cloud and expanding coverage to include Language processing (BERT) and SSD MobileNetv1 to Vision networks. We look forward to seeing additional submissions in the future, as well as any early indications of Qualcomm's plans for future Cloud AI platforms. When it comes to AI, we have no doubt that Qualcomm is in it to win it.

[4] MLPerf™ v1.1 Inference Closed Edge Power; Latency and Energy data derived from submissions:
Edge Device Qualcomm AEDK DM.2e 1.1-120, DM.2 1.1-121, Nvidia 1.1-111, 1.1-118

[5] MLPerf™ v1.1 Inference, Closed Datacenter, Closed Edge; offline, server performance data derived from submissions
Qualcomm 16x 1.1-056, 1.1-057 8x 1.1-058, 1.1-059, 5x 1.1-024, 1.1-025, 1x 1.1-022, 1.1-023

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder and Principal Analyst at Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with QTI funding and support. Although the document may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.