QUALCOMM CONTINUES TO DOMINATE POWER EFFICIENCY WITH MLPERF 2.1 INFERENCE

Company's recent design wins with leading server vendors indicates potentially high customer demand, especially in edge image processing.

Ever since Qualcomm announced its first-generation cloud edge AI processor, the Qualcomm Cloud AI 100, the company has been at the top of the leader board for power efficiency, a key customer requirement as the edge becomes part of a connected intelligent network. The latest benchmarks from MLCommons, an industry group of over 100 companies, demonstrates that the Qualcomm platform is still the most energy efficient AI accelerator in the industry for image processing.

In the latest results, Qualcomm partners Foxconn, Thundercomm, Inventec, Dell, HPE, and Lenovo all submitted leadership benchmarks, using the Qualcomm Cloud AI 100 "Standard" chip, which delivers 350 Trillion Operations Per Second (TOPS). The companies are, at least for now, targeting edge image processing where power consumption is critical, while Qualcomm Technologies submitted updated results for the pro SKU, targeting edge cloud inference processing.

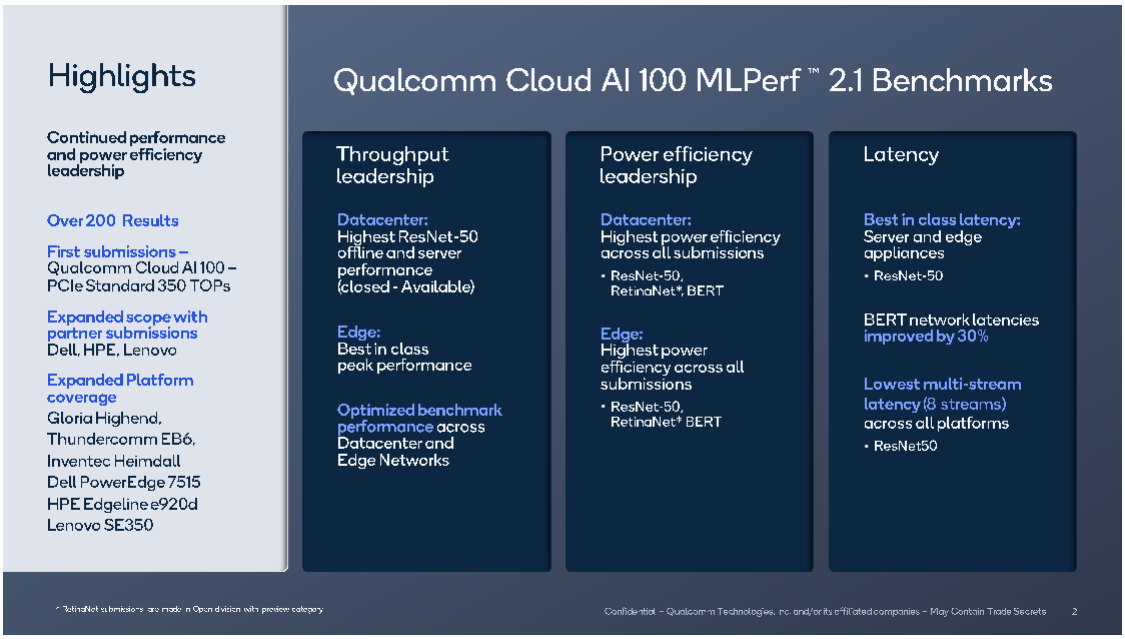

Figure 1: Highlights of submissions by Qualcomm and their partners to V2.1 MLPerf Inference.

The fact that both Gloria (Foxconn) and Inventec Heimdall, suppliers to cloud service companies, submitted results to MLCommons tells us that Qualcomm may be realizing traction in the Asian cloud market, while Dell, Lenovo, and HPE support indicates global interest in the Qualcomm part for datacenter and "Edge Clouds".

THE RESULTS

The current Cloud AI 100 demonstrates dramatically superior efficiency for image processing compared to both cloud and edge competitors. This makes sense as the heritage of the accelerator is the high-end Snapdragon mobile processor's Qualcomm AI Engine, which provides AI for mobile

handsets where imaging is the primary application. Nonetheless, the Qualcomm platform provides best-in-class performance efficiency for the BERT-99 model, used in natural language processing.

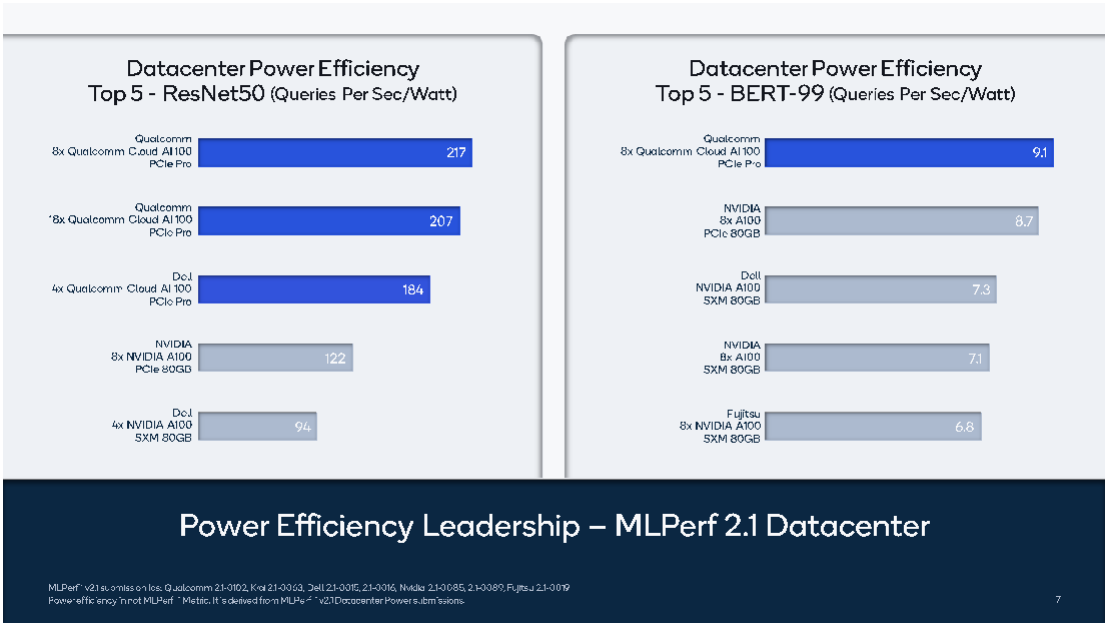

Figure 2: The Qualcomm Cloud AI 100 beats the NVIDIA A100, the industry gold standard for AI acceleration, in performance per watt.

Figure 3 shows how well Qualcomm Cloud AI 100 performs in edge image processing (ResNet50**)** versus the NVIDIA Jetson Orin processor. We suspect that Qualcomm will extend the next design for the Cloud AI family beyond image processing as the cloud edge begins to require language processing and recommendation engines.

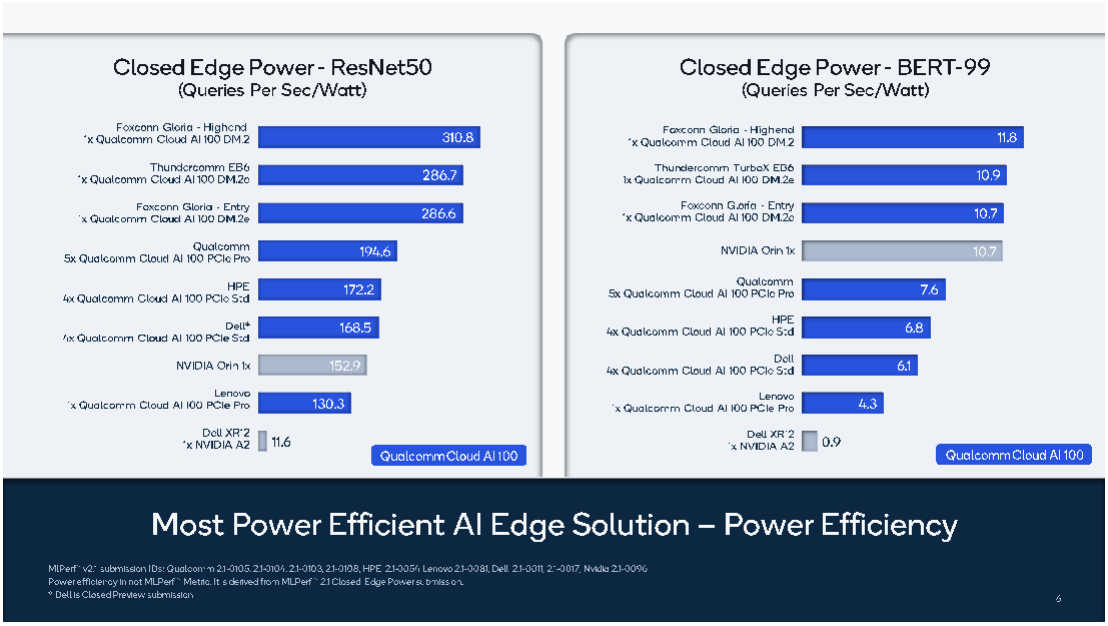

Figure 3: In image processing, the Qualcomm Cloud AI 100 doubles the performance per watt over NVIDIA Jetson Orin with the standard part, while in Language processing, Qualcomm still delivers marginally better efficiency.

Not surprisingly, Qualcomm's power efficiency does not come at the expense of high performance, unlike

many startups targeting this emerging market. As Figure 4 shows, a 5-card server, each consuming 75 watts, delivers nearly 50 percent more performance than a 2xNVIDIA A100, each of which consumes 300 watts. In an analysis of the potential economic benefits of this power efficiency, we estimate that a large data center could save 10's of millions of dollars a year in energy and capital costs by deploying the Qualcomm Cloud AI 100.

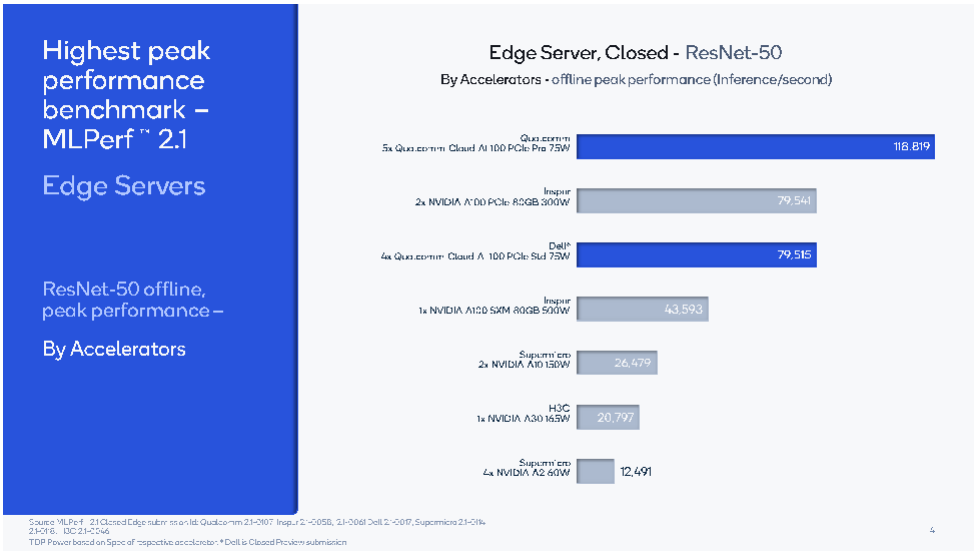

Figure 4: Qualcomm's power efficiency does not come at the expense of high performance.

CONCLUSIONS

Qualcomm is still the champion when it comes to power efficiency, but there are two potential issues they face. First, while NVIDIA's Hopper-based H100 showed the fastest inference performance submitted to MLCommons, that platform is not yet generally available, and NVIDIA has not submitted any power measurements. We suspect that the H100 may eclipse Qualcomm's star for efficiency, but we also suspect this may be a fleeting claim to fame, as we will probably see a second-generation part in a similar timeframe, or perhaps a few months later. Second, while the Qualcomm Cloud AI 100 has exceptional efficiency and performance for image processing, it does not blow NVIDIA's A100 away for NLP, and we have yet to see performance data for other models such as recommendation engines. Consequently, while an edge AI processor typically only requires image analysis, a large data center may choose to await more even model coverage the next generation could provide.

IMPORTANT INFORMATION ABOUT THIS PAPER

Authors: **Karl Freund, Cambrian-AI Research, LLC**

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.