

Qualcomm

Qualcomm[®] Cloud AI 100 Ultra

Large-Language Model (LLM) inference Performance
April 2024

Qualcomm Cloud AI 100 Ultra

LLM Inference Performance

Model	Input Length	Output Length	Avg Token Latency	Server throughput (tok/s) - 16x AI100
LLaMa 7B	128	128	< 10 ms	> 1500
LLaMa 7B	2048	2048	< 10 ms	> 1350
LLaMa 70B	128	128	< 35 ms	> 110
LLaMa 70B	2048	2048	< 35 ms	> 100

Latency Mode (Batch = 1)

Model	Input Length	Output Length	Server throughput (tok/s) - 16x AI100
LLaMa 7B	128	128	> 46000
LLaMa 7B	2048	2048	> 7300
LLaMa 70B	128	128	> 5700
LLaMa 70B	2048	2048	> 3500

Throughput Mode (Batch = 64)

- Fp16 compute, MXFP6 weights. Performance reported at server level with 16x AI 100 Ultra cards
- Utilizing advanced optimization techniques can result in additional performance increase of up to 4x

Qualcomm Cloud AI 100: Leading in all inference performance metrics with industry-leading advancements in performance density and performance-per-watt capabilities.

Thank you

Qualcomm

Follow us on: [in](#) [twitter](#) [instagram](#) [youtube](#) [facebook](#)

For more information, visit us at:

qualcomm.com & qualcomm.com/blog

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018-2023 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.