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:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog