

Qualcomm

# Qualcomm<sup>®</sup> Cloud AI 100

MLPerf™ v4.0

Inference Benchmarks

# Qualcomm Cloud AI 100 MLPerf™ 4.0 Benchmarks

## Highlights

### Industry leading Performance and power efficiency

#### Introducing Qualcomm Cloud AI 100 Ultra -

#### Power optimized GenAI Inference accelerators to MLPerf™

- Early - Closed Preview Submission for Qualcomm Cloud AI 100 Ultra AI Inference Accelerators
- Early Preview submissions from Dell HPE and Lenovo

## Industry-leading power efficiency

### Highest power efficiency\*

- ResNet-50 : up to 275 Inference/Sec/Watt
- RetinaNet : up to 5.2 inference/Sec/Watt

### Best In Class power efficiency\*

- BERT-99 : up to 10.18 inference/Sec/Watt
- Stable Diffusion XL : Low power to generate every image

## Industry-leading throughput

### Highest Offline performance<sup>1</sup>

- ResNet-50 : > 902K Inference/Sec for 16x Qualcomm Cloud AI 100 Datacenter server

### Best in class Performance across all submitted platforms in datacenter and edge categories

### Benchmarks on Public Qualcomm Cloud AI 100 based AI accelerators instances

- AWS DL2Q Instances
- Cirrascale Cloud Instances

\* Closed Preview Submission for Qualcomm Cloud AI 100 Ultra  
MLPerf™ v4.0 Submission IDs: 4.0-0085, 4.0-0086, 4.0-0066

# Power Efficient Datacenter AI solutions

MLPerf™ 4.0 – Closed  
Preview Division

## 16x Qualcomm Cloud AI 100 Ultra\*

Sub ID	Network	Power Efficiency
4.0-0085	ResNet-50	275.0 Perf/Watt
4.0-0085	RetinaNet	5.2 Perf/Watt
4.0-0085	BERT - 99	10.2 Perf/Watt

Best In Class  
Performance  
Datacenter AI  
solutions

MLPerf™ 4.0 – Closed  
Preview Division

## 16x Qualcomm Cloud AI 100 Ultra\*

Sub-ID	Network	Performance
4.0-0086	ResNet-50	902,482 Inference/Sec
4.0-0086	RetinaNet	15,477 Inference/Sec
4.0-0086	BERT - 99	30,966 Inference/Sec

Power Efficient  
Edge AI solutions

MLPerf™ 4.0 – Closed  
Preview Division

## 2x Qualcomm Cloud AI 100 Ultra\*

Sub ID	Network	Power Efficiency
4.0-0087	ResNet-50	213 Perf/Watt
4.0-0087	RetinaNet	4.1 Perf/Watt
4.0-0087	BERT - 99	8.0 Perf/Watt

Best In Class  
Performance  
Edge AI solutions

MLPerf™ 4.0 – Closed  
Preview Division

## 2x Qualcomm Cloud AI 100 Ultra\*

Sub-ID	Network	Performance	Remarks
4.0-0088	ResNet-50	122,566 Inference/Sec	
4.0-0088	RetinaNet	2,079 Inference/Sec	
4.0-0088	BERT - 99	4,147 Inference/Sec	
4.0-0088	Stable Diffusion XL** Base 1.0	0.36 Images/Sec	Leads to ~ 3 images/Sec for 16x Ultra based Server

\*Qualcomm Cloud AI 100 Ultra - Closed Preview Submission

\*\*SDXL 3 images/Sec performance for 16x Qualcomm Cloud AI 100 Ultra is extrapolated from Edge submission 4.0-0088

Power Efficient AI  
Edge solutions

MLPerf™ 4.0 – Closed  
Available Division

## 8x Qualcomm Cloud AI 100 Pro

Sub ID	Network	Power Efficiency
4.0-0066	ResNet-50	213.8 Perf/Watt
4.0-0066	RetinaNet	4.1 Perf/Watt
4.0-0066	BERT - 99	8.2 Perf/Watt
4.0-0066	Stable diffusion XL	0.0011 Images/Watt

Best In Class  
Performance - AI  
Edge solutions

MLPerf™ 4.0 – Closed  
Available Division

## 8x Qualcomm Cloud AI 100 Pro

Sub-ID	Network	Performance
4.0-0066	ResNet-50	188,415 Inference/Sec
4.0-0066	RetinaNet	2,516 Inference/Sec
4.0-0066	BERT - 99	6,313 Inference/Sec
4.0-0066	SDXL Base 1.0	0.61 Images/Sec



# Cloud AI 100 – Cloud Instances Benchmarks

MLPerf™ 4.0 – Closed Available Division

## 8x Qualcomm Cloud AI 100 Pro/standard

Sub-ID	Network	Performance for AWS dl2q.24xlarge instance (8x Cloud AI 100 Standard)	Performance for Cirrascale AI 100 Quad Instance (4x Cloud AI 100 Pro)
4.0-096	ResNet-50	157,977 Inference/Sec	
4.0-007	RetinaNet	2,494 Inference/Sec	
4.0-005	BERT - 99		3,150 Inference/Sec

# Thank you

**Qualcomm**

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

For more information, visit us at:

[qualcomm.com](http://qualcomm.com) & [qualcomm.com/blog](http://qualcomm.com/blog)

All data and information contained in or disclosed by this document is confidential and proprietary information of Qualcomm Technologies, Inc. and/or its affiliated companies and all rights therein are expressly reserved. By accepting this material the recipient agrees that this material and the information contained therein will not be used, copied, reproduced in whole or in part, nor its contents revealed in any manner to others without the express written permission of Qualcomm Technologies, Inc. Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018-2022 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.