

Qualcomm

Qualcomm® Cloud AI 100

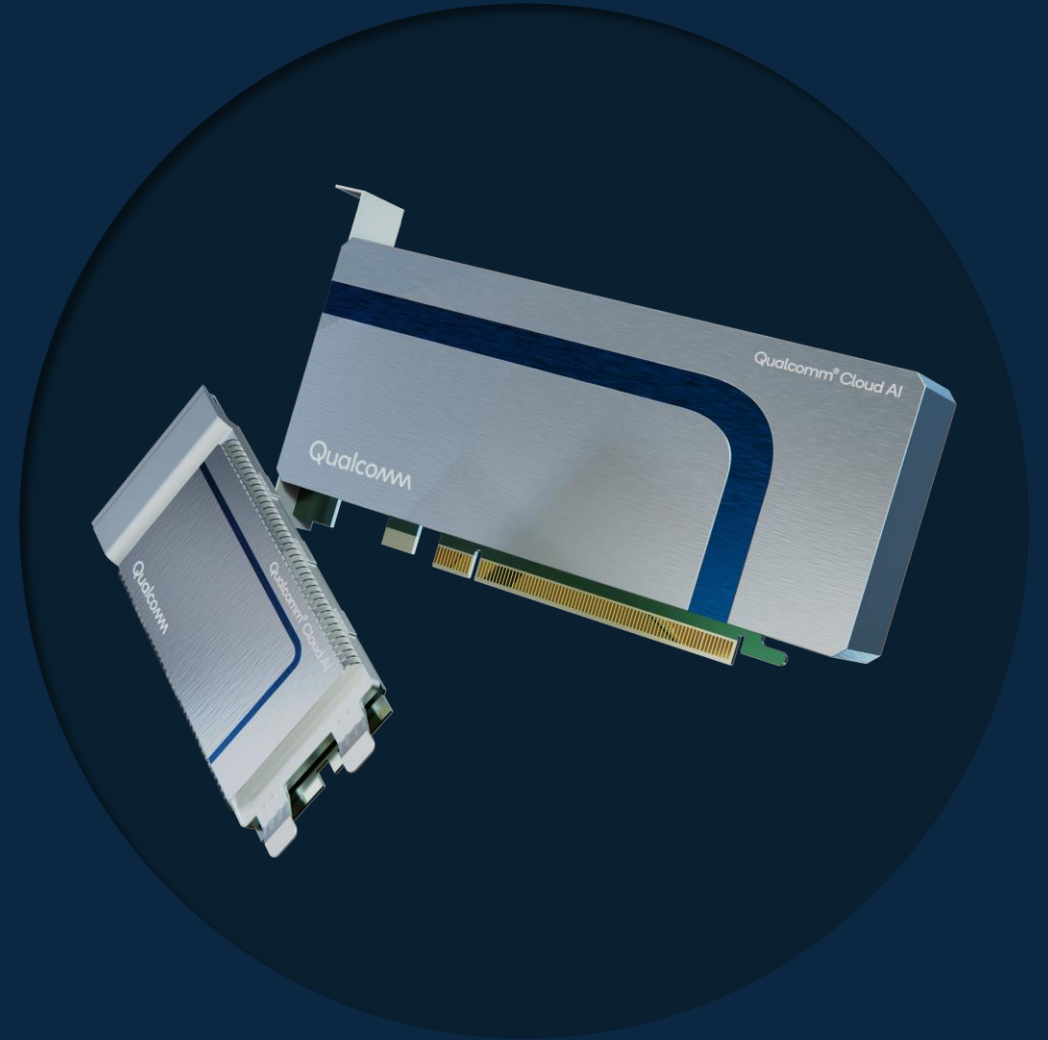
MLPerf™ v3.1

Inference Benchmarks

Nitin Vairagare

Director of Product Management

Qualcomm Technologies, Inc.



Qualcomm Cloud AI 100 MLPerf™ 3.1 Benchmarks

Industry leading power efficiency,
Best in Class Performance

160+ Results, 12 Platforms

First submission –

Inference Over Network – RetinaNet

Partner Platform coverage

HPE e920d, DL 385

Dell XR4520c

Lenovo SR665v1, SE450

Industry leading power efficiency

Datacenter:

Highest power efficiency
across platforms (QPS/W)

- ResNet-50 **61% higher**
- RetinaNet **41% higher**

over nearest competitor

Edge:

Highest power efficiency
across platforms (QPS/W)

- ResNet-50 **312 QPS/Watt**
- RetinaNet **5.8 QPS/Watt**
- BERT **12.1 QPS/Watt**

Latency

Low Single stream latency
at lowest power for Edge
Platforms:

- ResNet-50 - **9.53 Milli Joules per stream**

Lowest multi-stream
latency (8 streams) at
lowest power across all
platforms

- ResNet-50 - **38 Milli joules for 8 streams**

Industry-leading throughput

Highest Offline
performance across all
Edge platforms

Best in class Performance
across all platforms
for Datacenter

Optimized benchmark
performance across
Datacenter and Edge
Networks

- RetinaNet up to **12%**

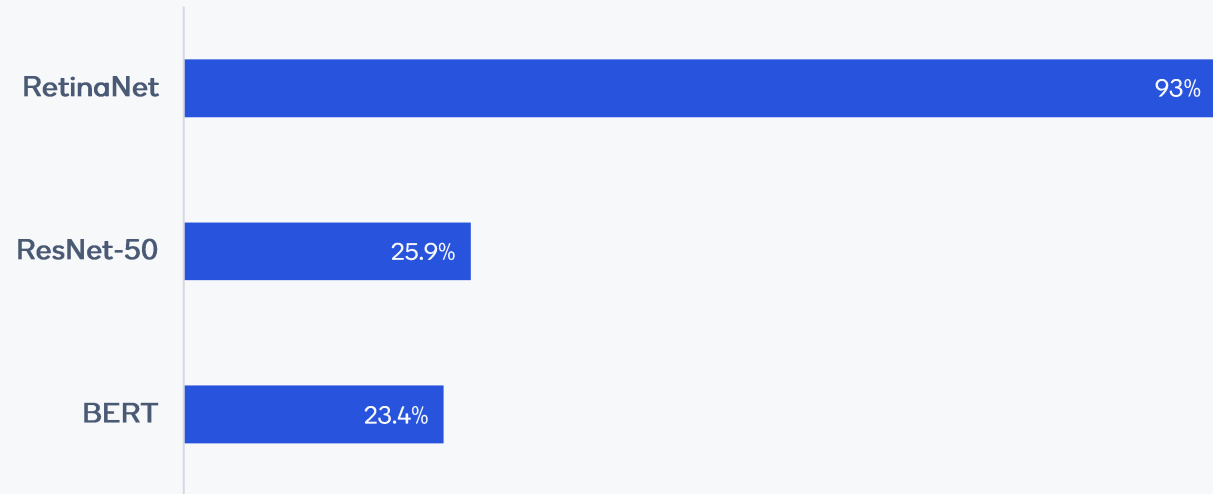
Proven Inference over Network (BERT, RetinaNet)
– Network Division Performance same as Closed Division

Qualcomm Cloud AI 100

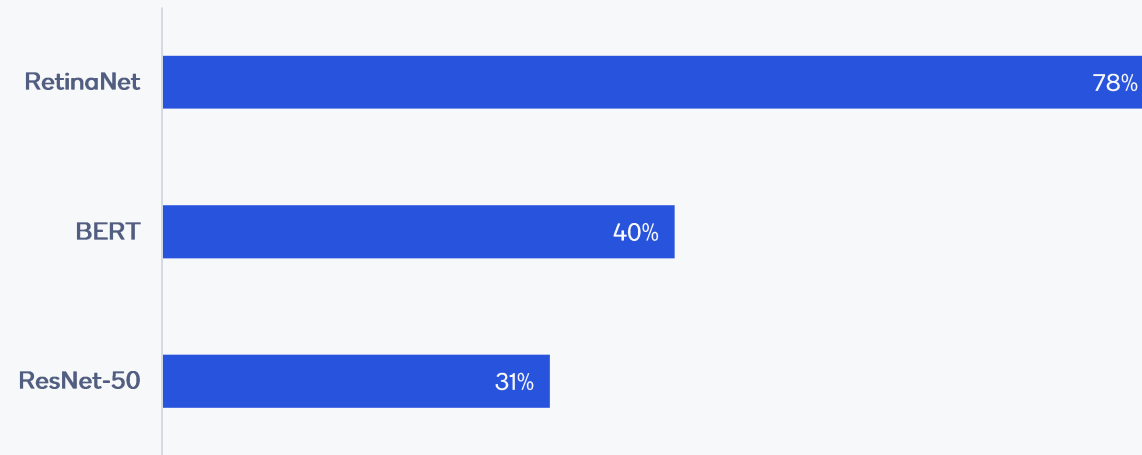
Continued
Performance and
Power efficiency
improvements via
Software optimizations

Since first MLPerf™
submission

Performance (QPS) Optimizations



Power Efficiency Optimizations



Most Power Efficient AI solutions

Computer vision

Across Datacenter and
Edge

MLPerf™ 3.1 - Closed
Division

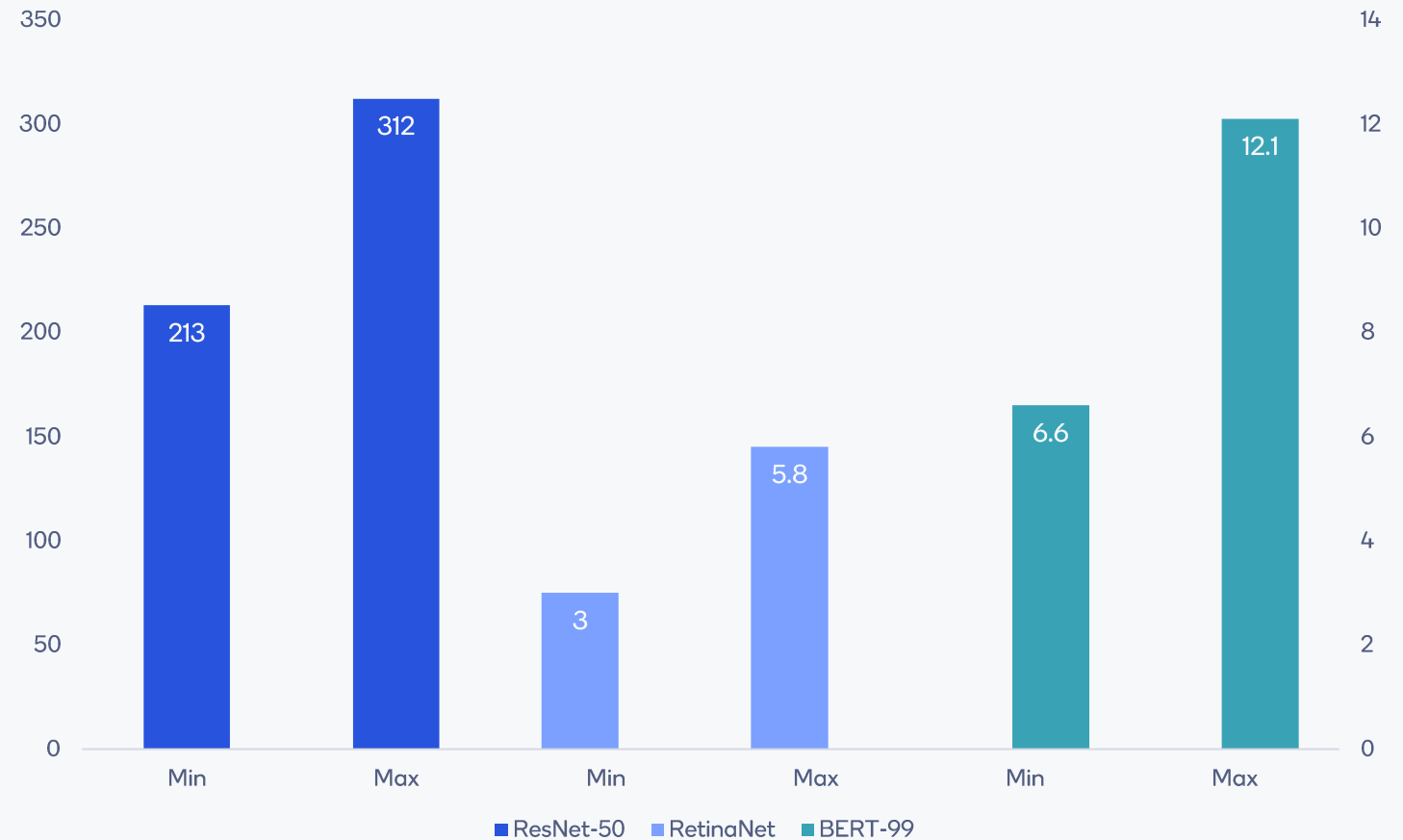
Qualcomm Cloud AI 100 SKU

Qualcomm Cloud AI 100 DM.2e

Qualcomm Cloud AI 100 PCIe Standard

Qualcomm Cloud AI 100 PCIe Pro

Qualcomm Cloud AI 100 Architecture Power Efficiency (QPS/Watt) Range across Cloud AI 100 SKU

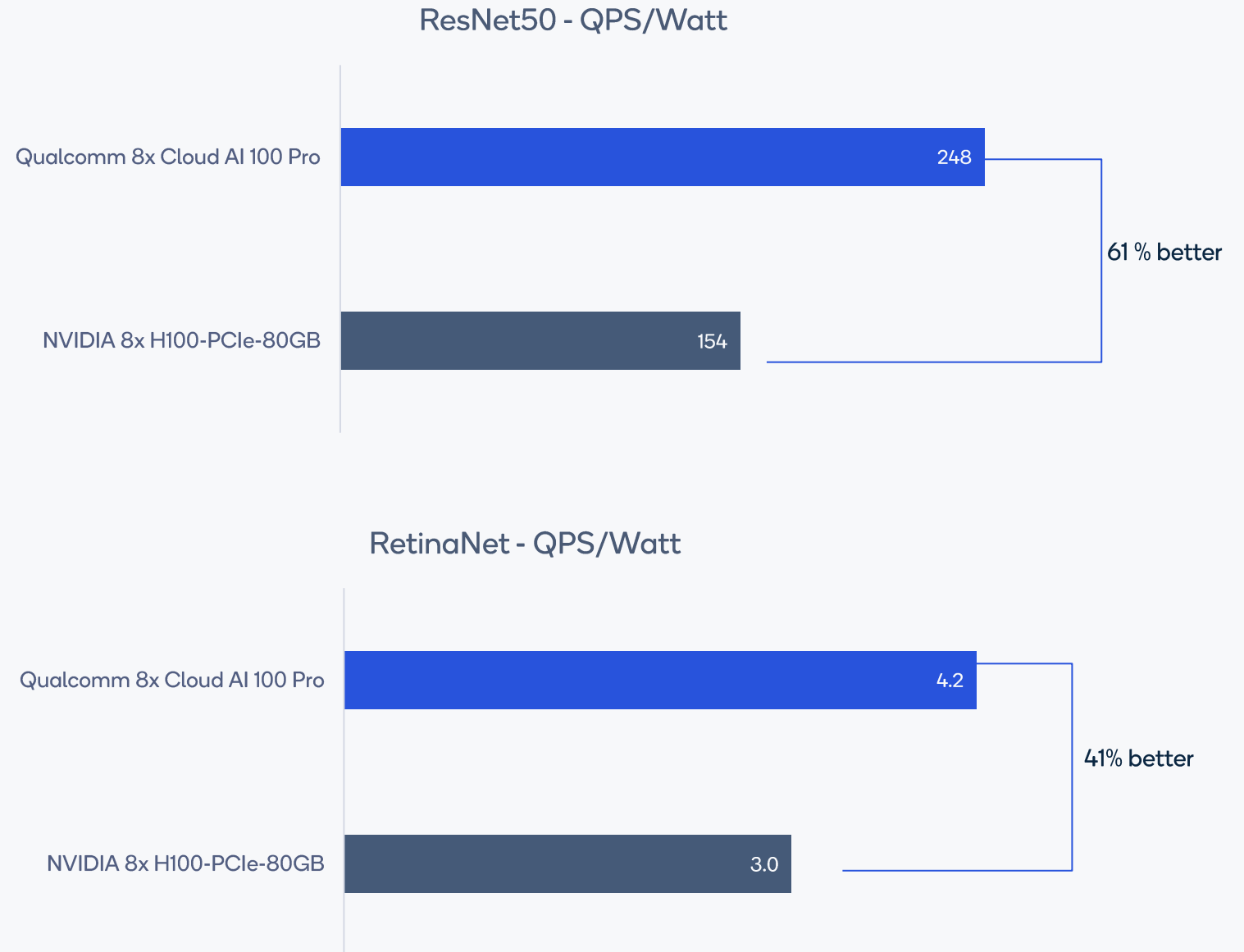


Most Power Efficient AI solutions Computer vision

Qualcomm Cloud AI 100 Vs Nearest Competitor "Gen + 1" Accelerator

Datacenter MLPerf™ 3.1 - Closed Division

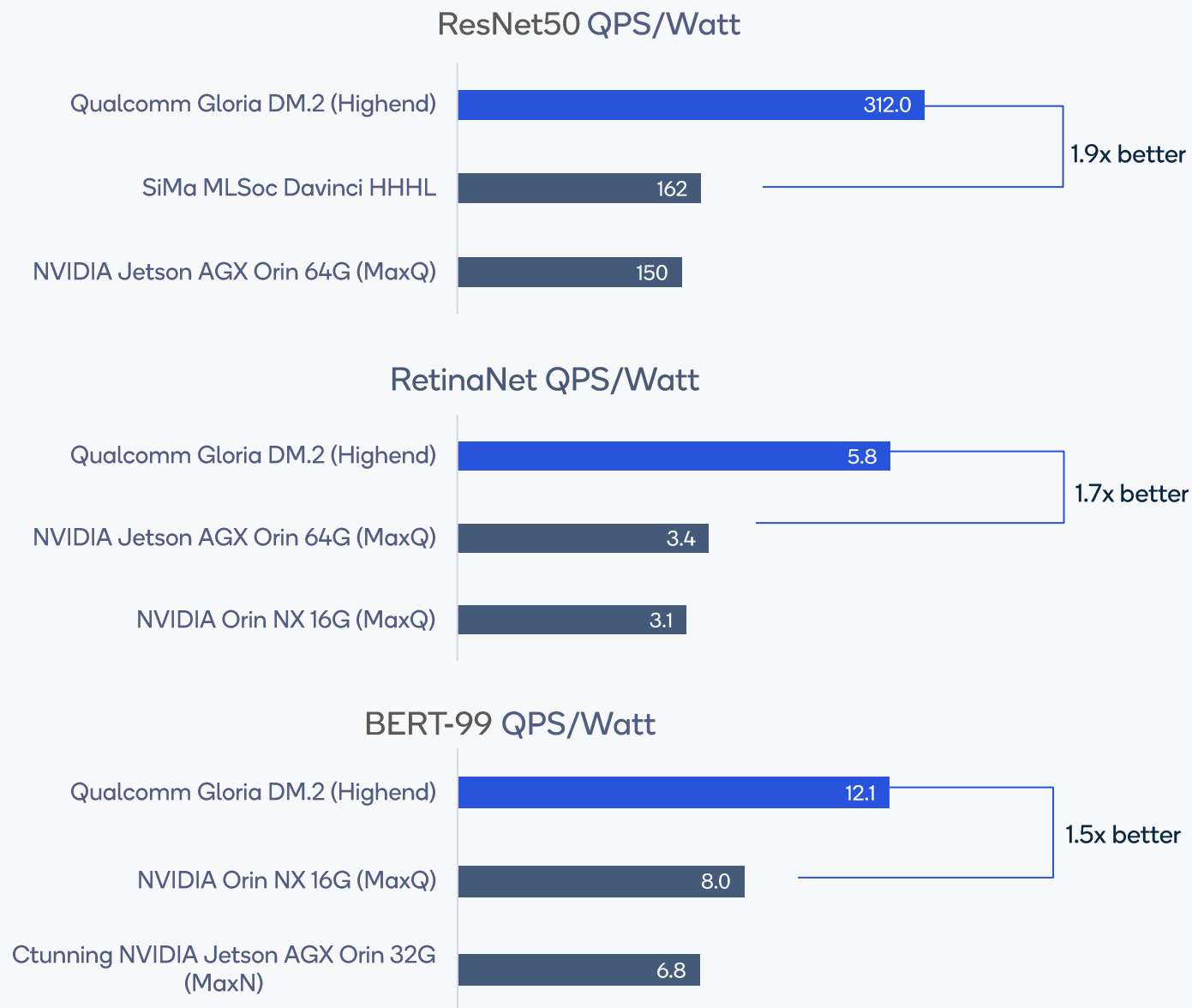
Datacenter Power Efficiency



Most Power Efficient AI solutions Computer vision

Edge Appliances MLPerf™ 3.1 - Closed Division

Power Efficiency - Edge Appliances

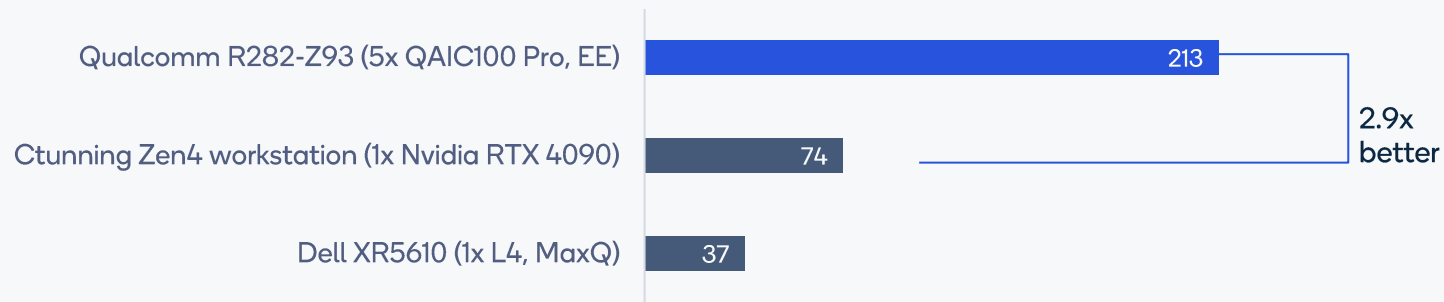


Most Power Efficient AI solutions Computer vision

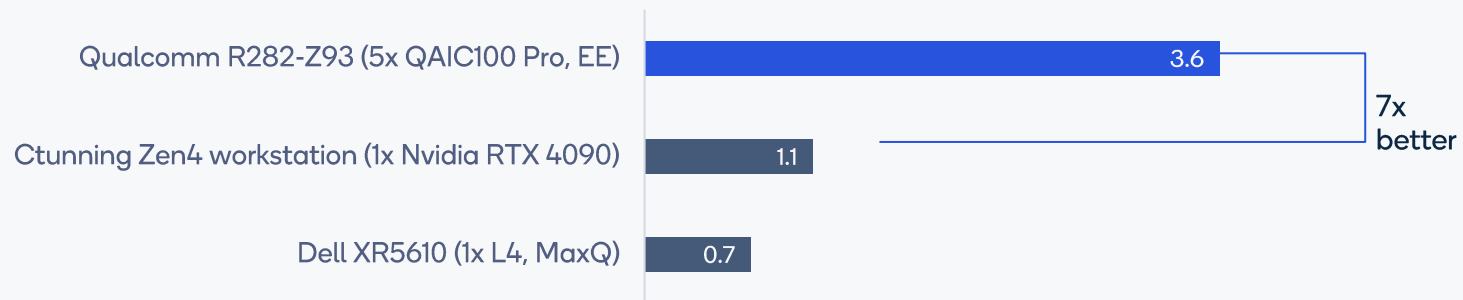
Edge Servers MLPerf™ 3.1 - Closed Division

Power Efficiency - Edge Servers

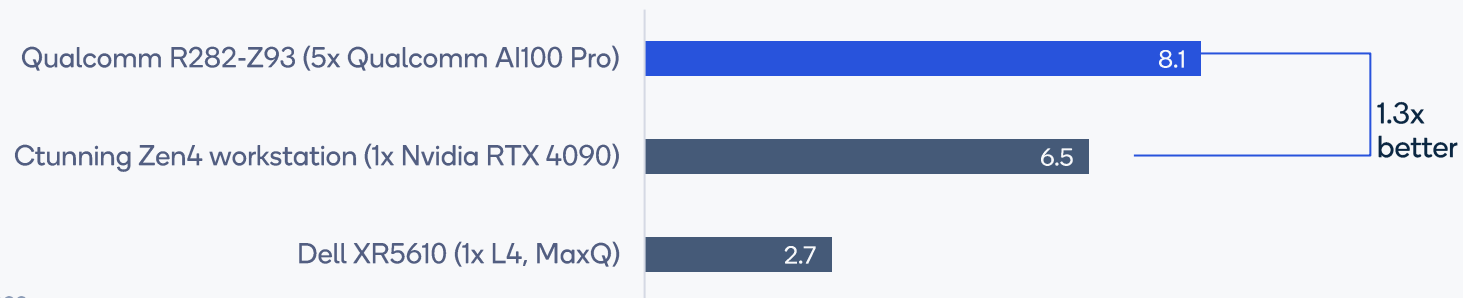
ResNet-50 QPS/Watt



RetinaNet QPS/Watt

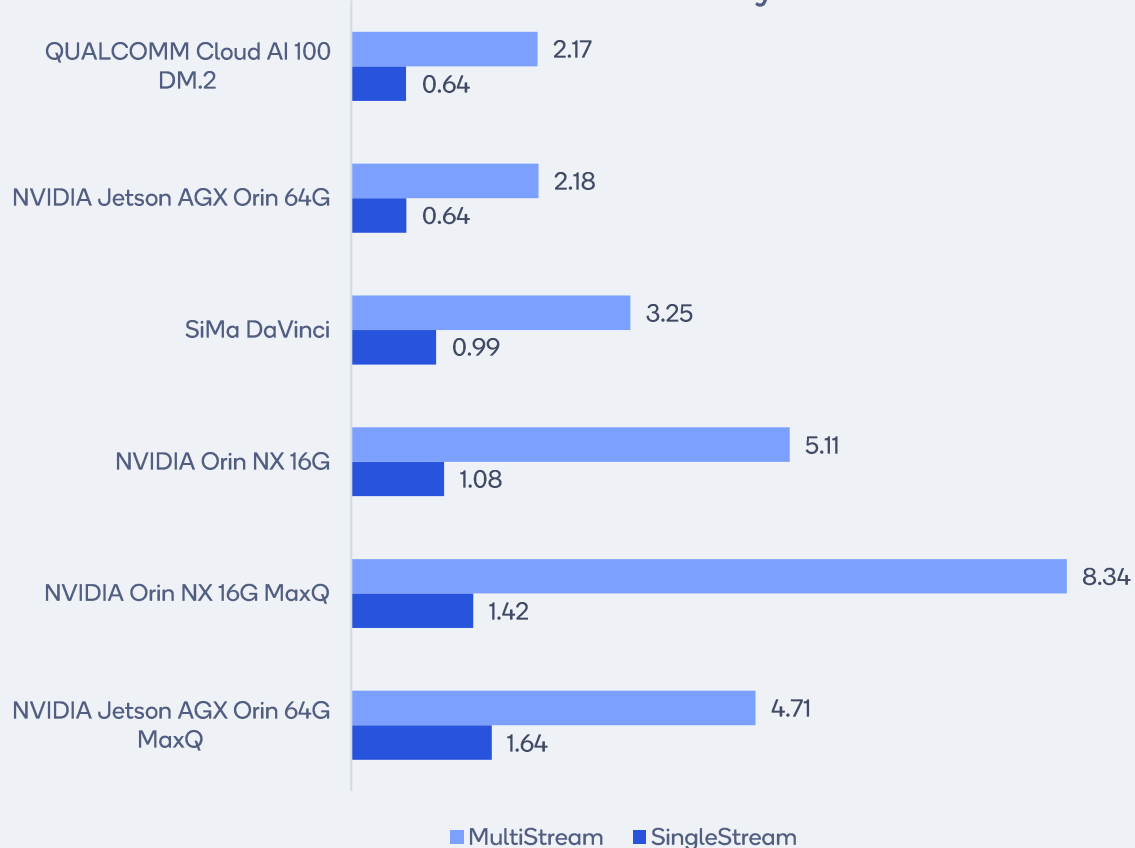


BERT-99 - QPS/Watt



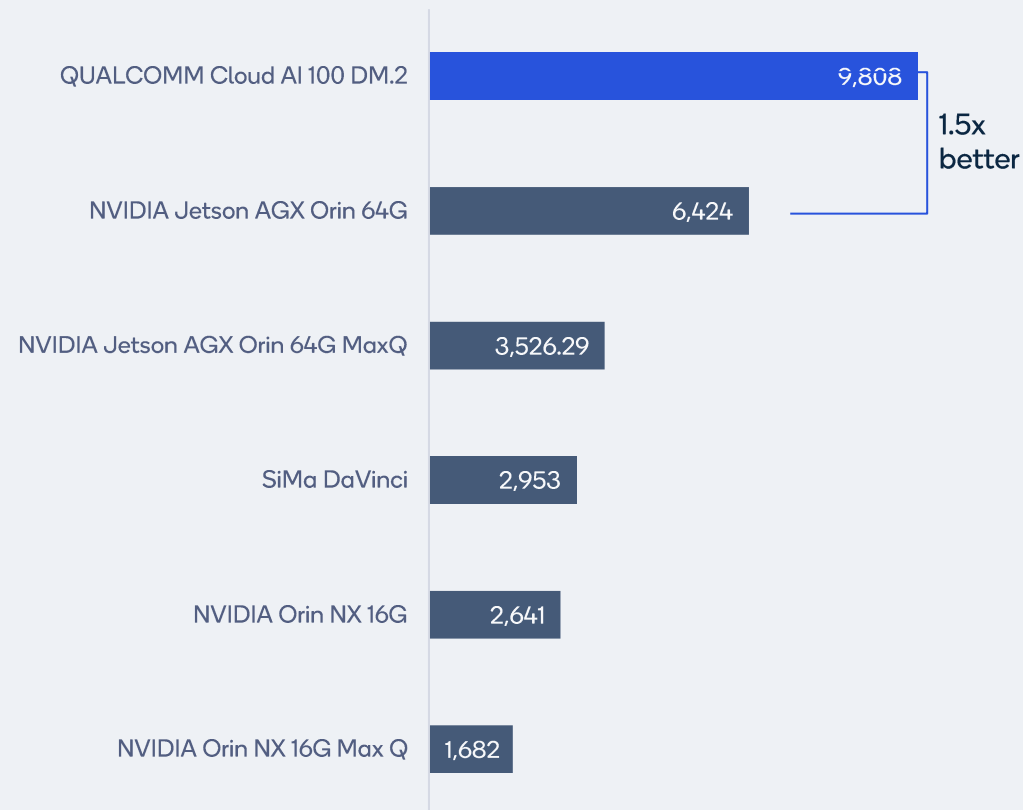
Edge Devices – Latencies (ms)

ResNet-50 – Lower latency is better



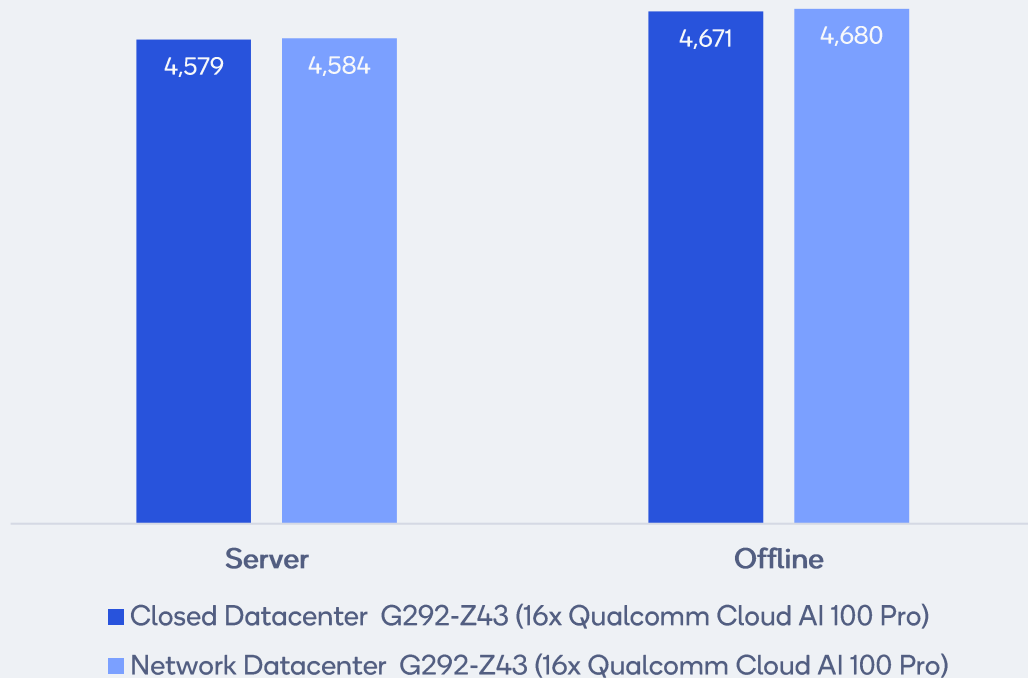
Edge Devices - Offline Performance

ResNet-50– Samples/Sec

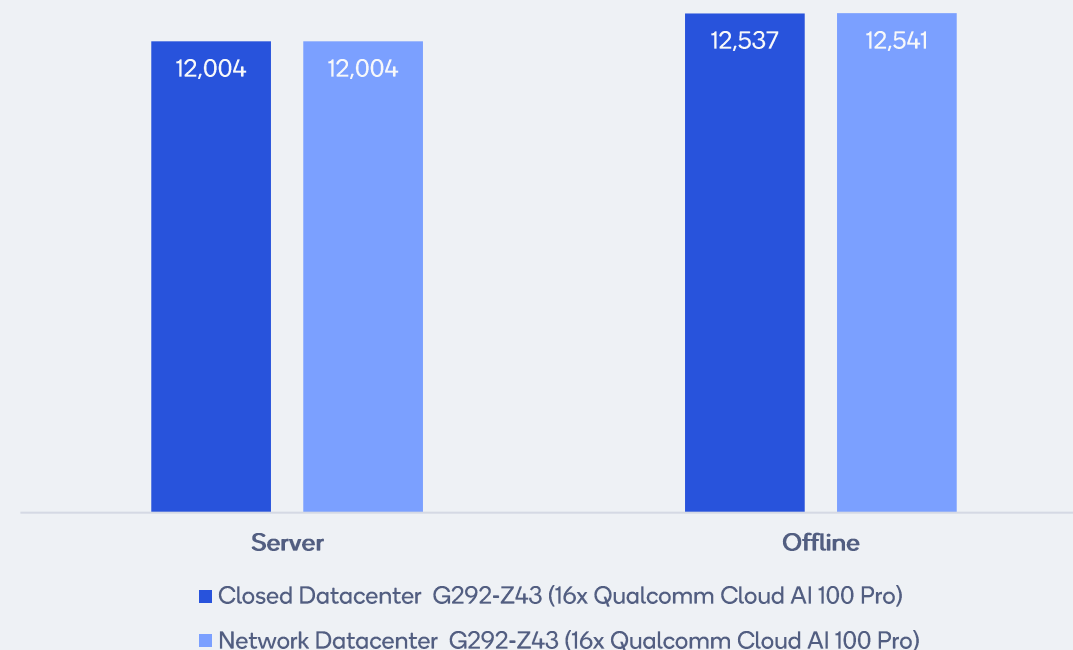


Qualcomm Cloud AI 100 Edge Devices Highest offline Performance, Lowest Latencies

Closed Vs Network Datacenter Qualcomm Cloud AI 100 PCIe Pro – RetinaNet



Closed Vs Network Datacenter Qualcomm Cloud AI 100 PCIe Standard – BERT-99



Qualcomm Cloud AI 100 Inference Over Network
Network Datacenter performance (QPS) as good as Closed Datacenter

Thank you



Follow us on:     

For more information, visit us at:

qualcomm.com & 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.

©2018-2023 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.