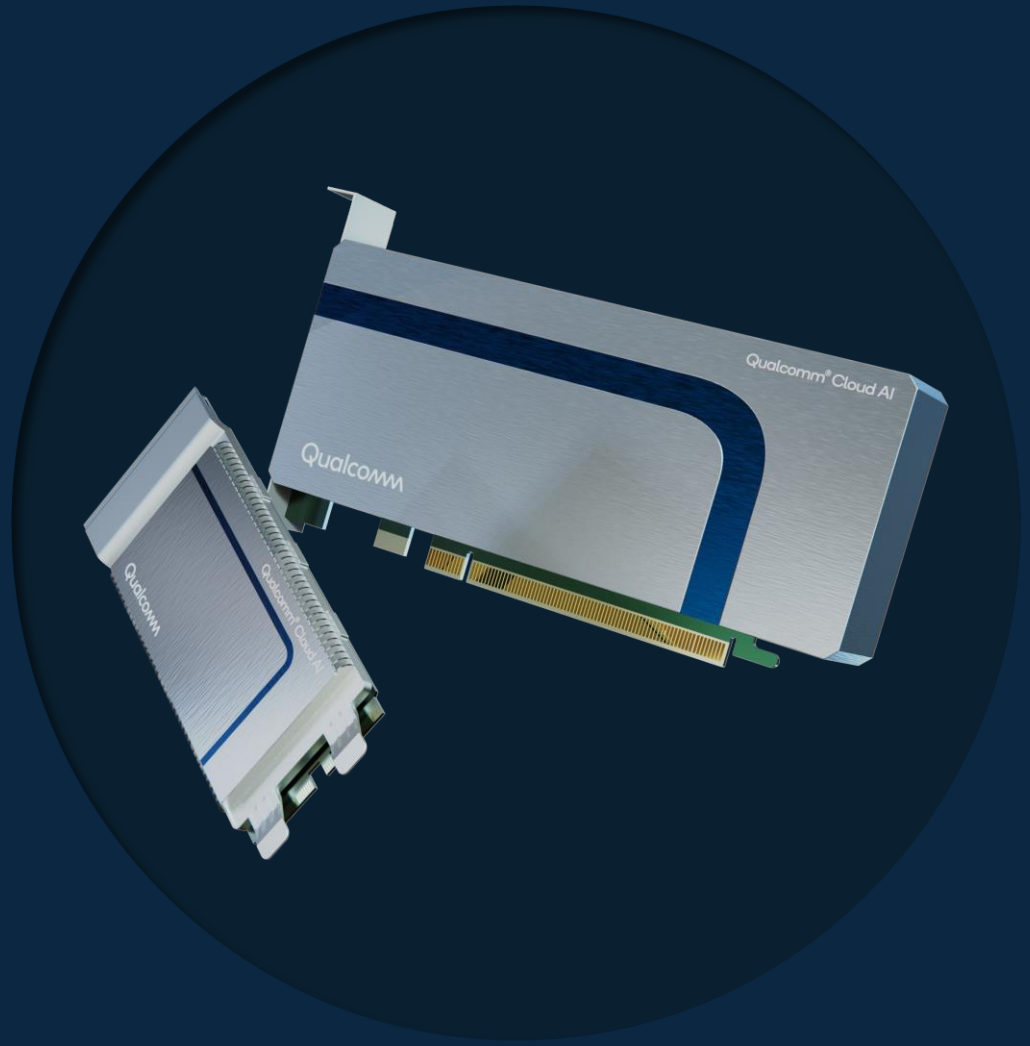
Qualcomm

Qualcomm® Cloud AI 100

MLPerf™ v2.1

Inference Benchmarks

John Kehrli, Sr Director of Product Management at Qualcomm Technologies Inc.



Highlights

Continued performance and power efficiency leadership

Over 200 Results

First submissions –
Qualcomm Cloud AI 100 – PCIe Standard 350 TOPs

Expanded scope with valuable submissions
Dell, HPE, Lenovo

Expanded Platform coverage
Gloria Highend,
Thundercomm EB6,
Inventec Heimdall
Dell PowerEdge 7515
HPE Edgeline e920d
Lenovo SE350

Qualcomm® Cloud AI 100 with MLPerf ™ v2.1 Benchmarks

Throughput leadership

Datacenter:
Highest ResNet-50 offline and server performance (closed - Available)

Edge:
Best in class peak performance

Optimized benchmark performance across Datacenter and Edge Networks

Power efficiency leadership

Datacenter:
Highest power efficiency across all submissions

- ResNet-50, RetinaNet*, BERT

Edge:
Highest power efficiency across all submissions

- ResNet-50, RetinaNet* BERT

Latency

Best in class latency:
Server and edge appliances

- ResNet-50

BERT network latencies **improved by 30%**

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

- ResNet50

- RetinaNet submissions are made in Open division with preview category

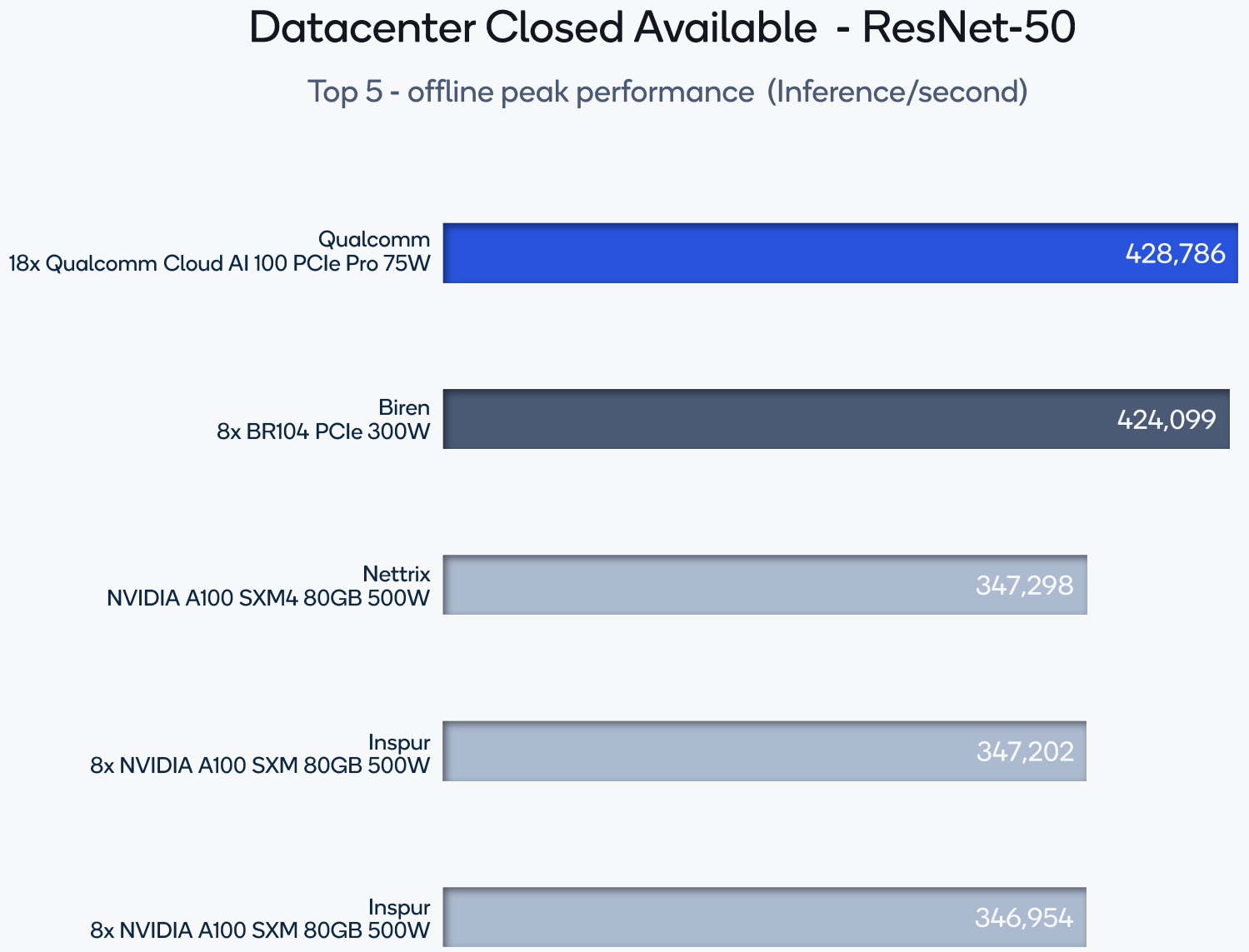
Qualcomm Cloud AI 100 is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

Highest peak performance benchmark – MLPerf ™ 2.1

Datacenter

Closed Available

ResNet-50 offline, peak performance –



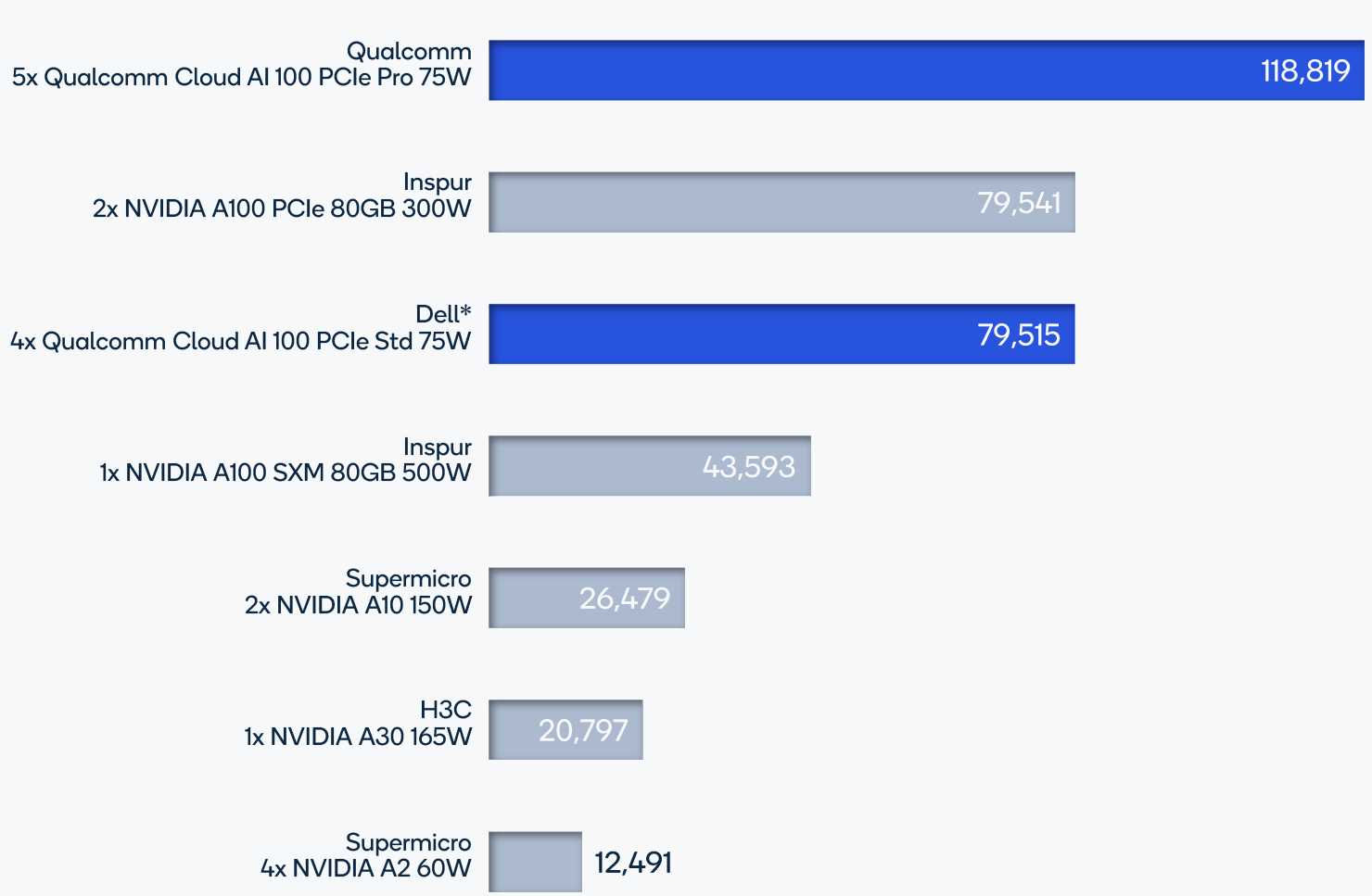
MLPerf™ 2.1 Closed Datacenter, Submission ID: Qualcomm 2.1-0101, Biren 2.1-0010, Nettrix 2.1-0097, Inspur 2.1-0057, 2.1-0056
TDP Power based on Spec of respective accelerator

Highest peak performance benchmark – MLPerf ™ 2.1

Edge Servers

ResNet-50 offline, peak performance –

By Accelerators



Source MLPerf™ 2.1 Closed Edge submission Id: Qualcomm 2.1-0107, Inspur 2.1-0058, 2.1-0061 Dell 2.1-0017, Supermicro 2.1-0114 2.1-0118, H3C 2.1-0046

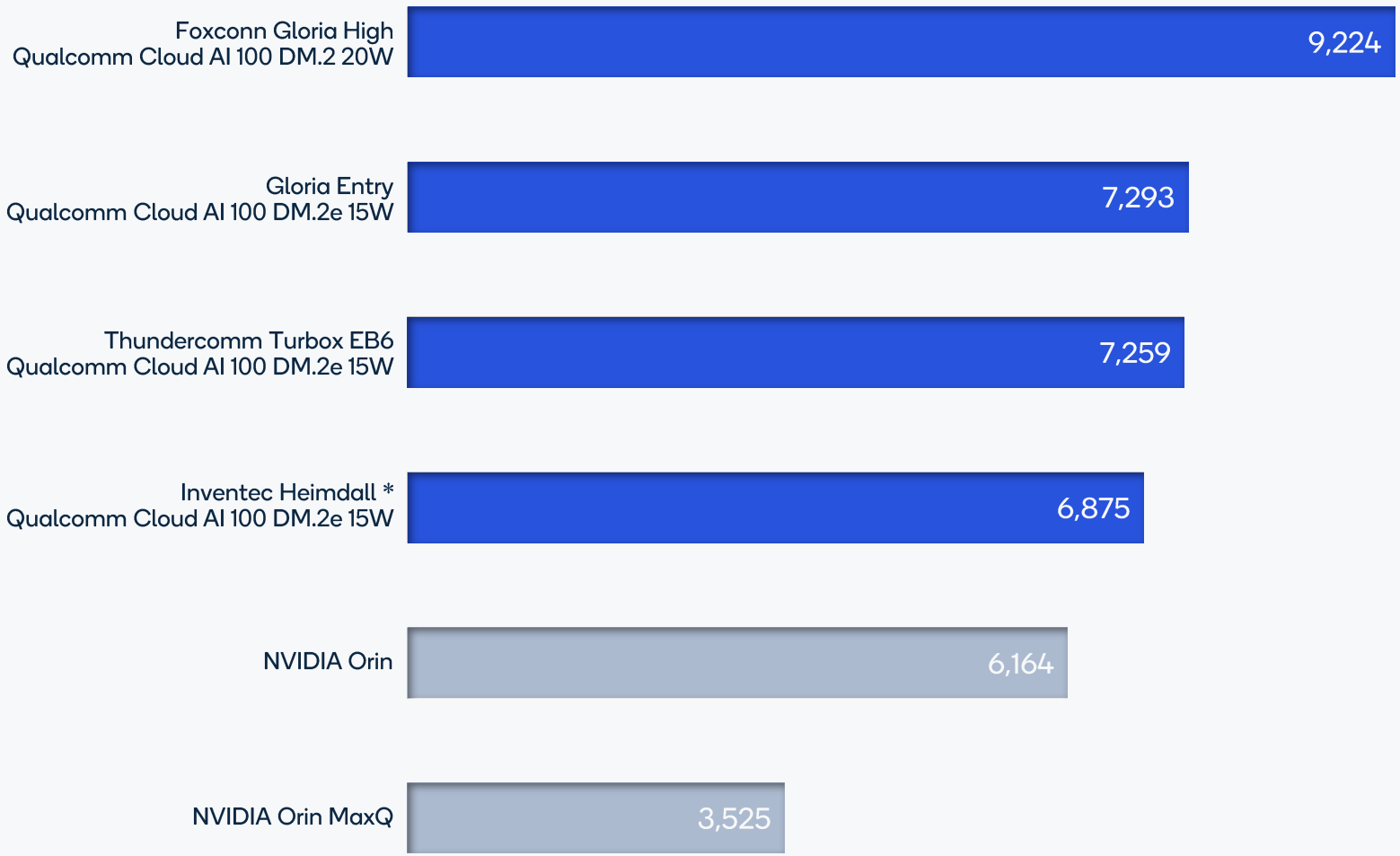
TDP Power based on Spec of respective accelerator. * Dell is Closed Preview submission

Highest peak performance benchmark – MLPerf ™ 2.1

Edge Appliance

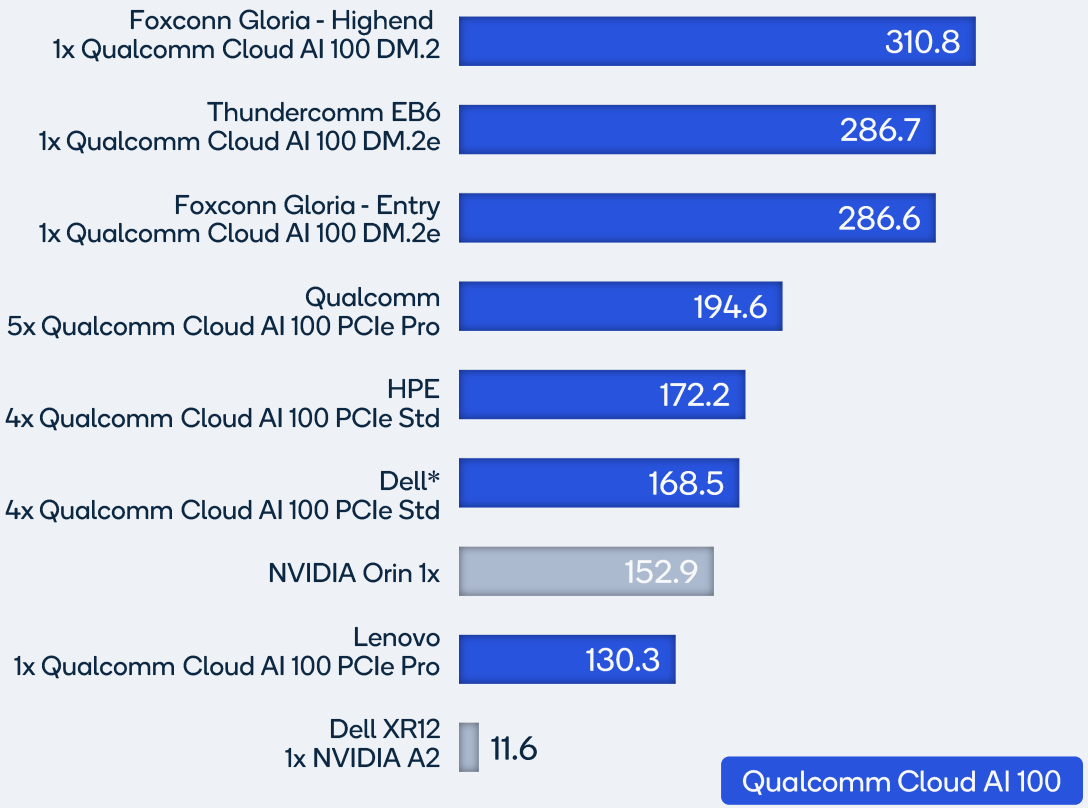
ResNet-50 offline, peak performance –

By Accelerators

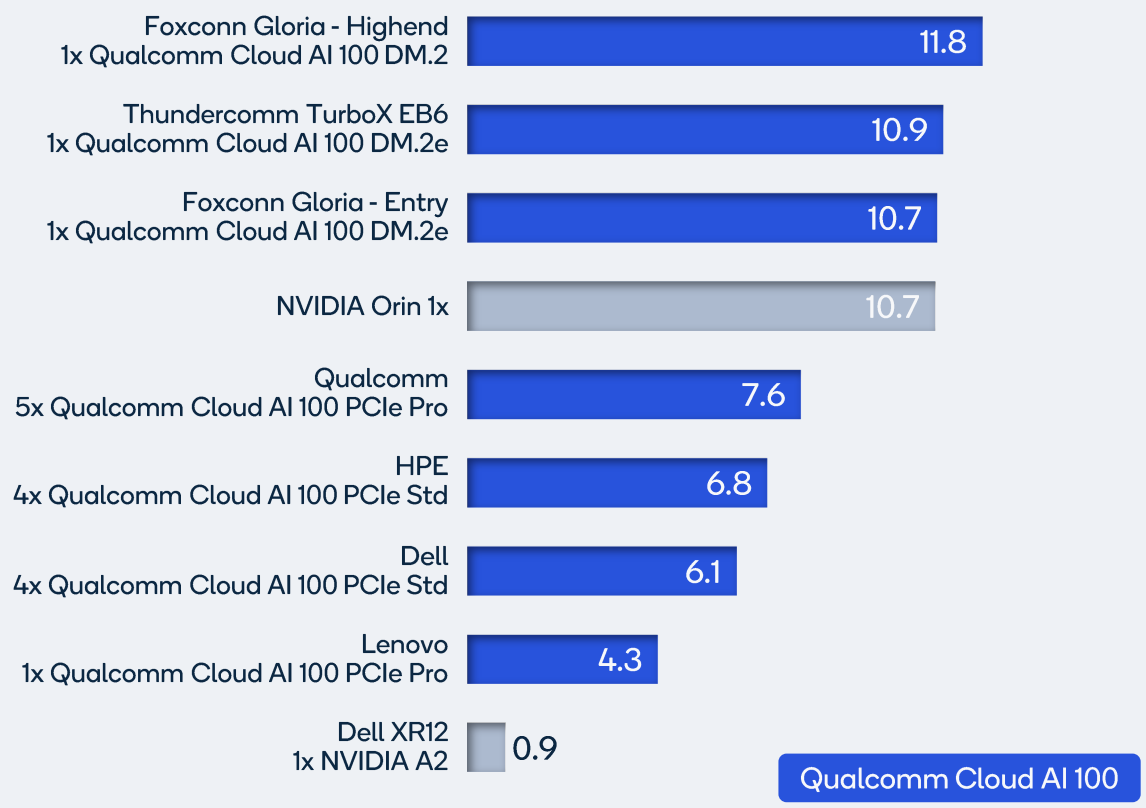


MLPerf ™ 2.1 Closed Datacenter Submission ID: Qualcomm 2.1-0105, 2.1-0104, 2.1-0103, 2.1-0123 Nvidia 2.1-0095, 2.1-0096
*Inventec Heimdall – Closed Preview

Closed Edge Power - ResNet50
(Queries Per Sec/Watt)



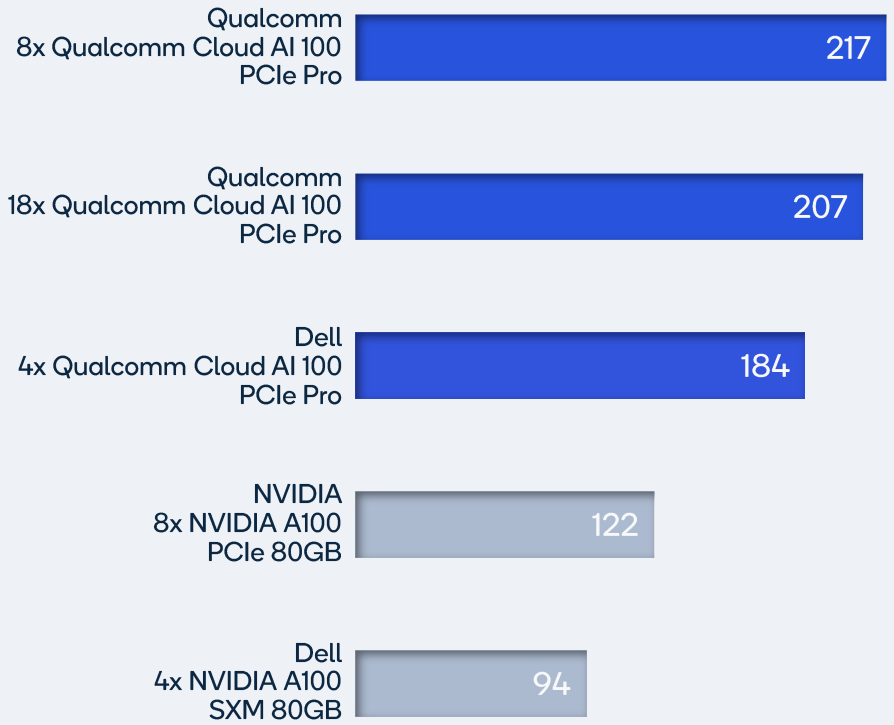
Closed Edge Power - BERT-99
(Queries Per Sec/Watt)



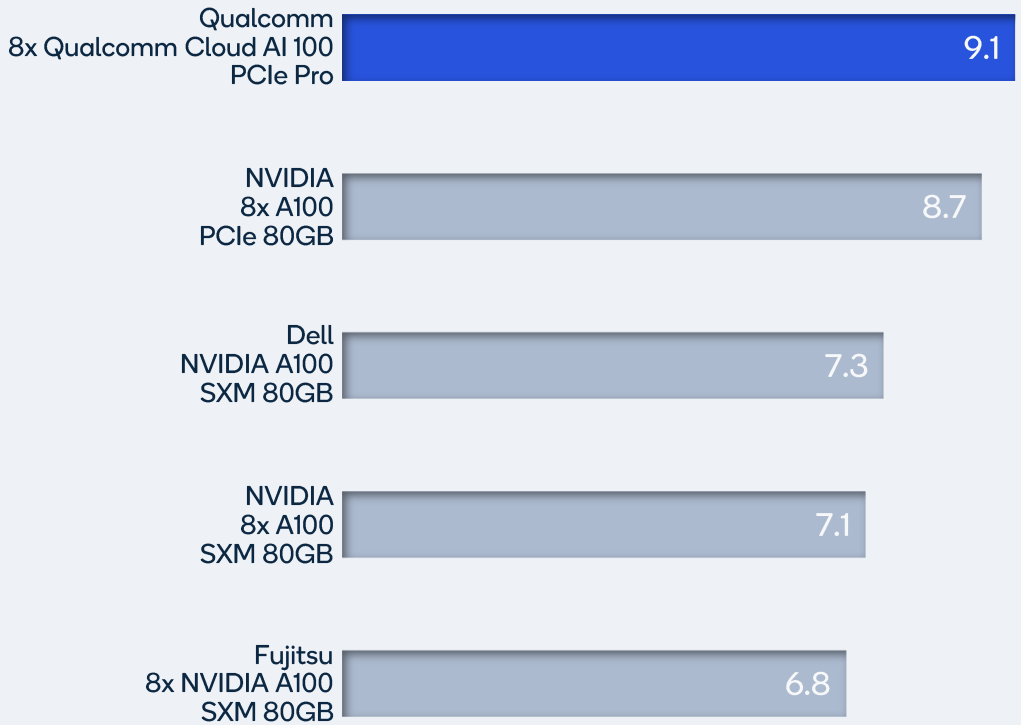
Most Power Efficient AI Edge Solution – Power Efficiency

MLPerf™ v2.1 submission IDs: Qualcomm 2.1-0105, 2.1-0104, 2.1-0103, 2.1-0108, HPE 2.1-0054 Lenovo 2.1-0081, Dell 2.1-0011, 2.1-0017, Nvidia 2.1-0096
Power efficiency in not MLPerf™ Metric. It is derived from MLPerf™ 2.1 Closed Edge Power submission.
* Dell is Closed Preview submission

Datacenter Power Efficiency
Top 5 - ResNet50 (Queries Per Sec/Watt)



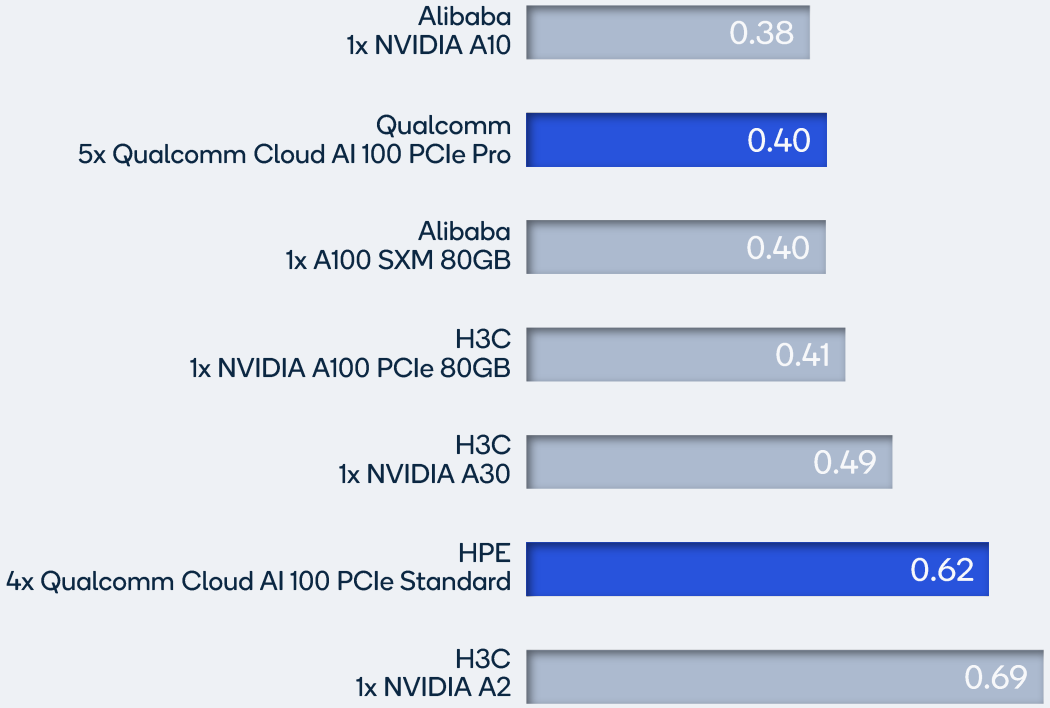
Datacenter Power Efficiency
Top 5 - BERT-99 (Queries Per Sec/Watt)



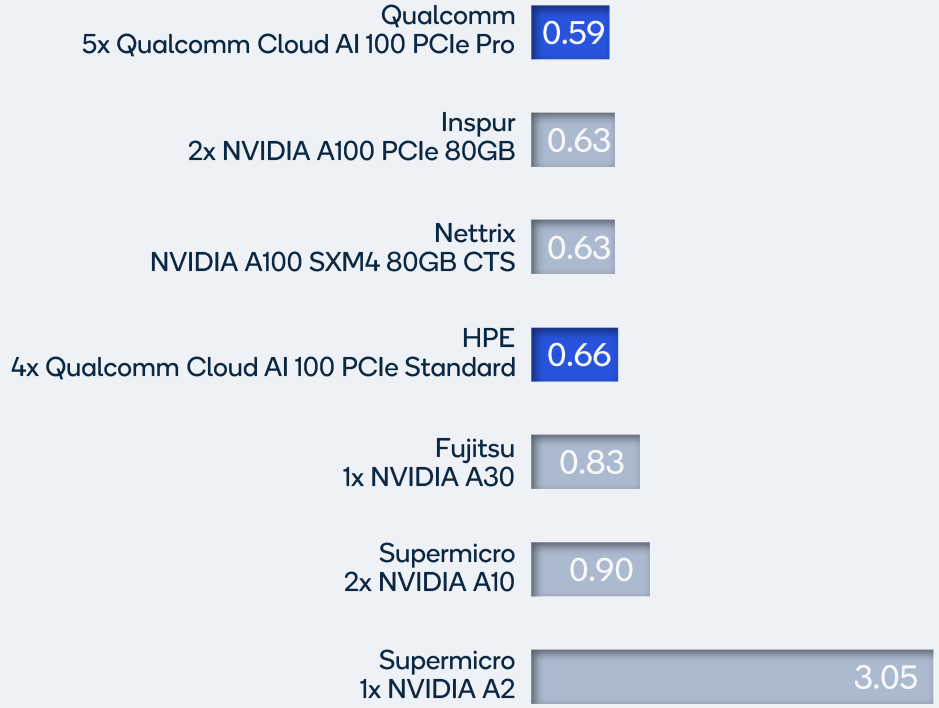
Power Efficiency Leadership – MLPerf 2.1 Datacenter

MLPerf™ v2.1 submission Ids: Qualcomm 2.1-0102, Krai 2.1-0063, Dell 2.1-0015, 2.1-0016, Nvidia 2.1-0085, 2.1-0089, Fujitsu 2.1-0019
Power efficiency in not MLPerf ™ Metric. It is derived from MLPerf ™ v2.1 Datacenter Power submissions.

Edge ResNet-50 Single stream Latencies (ms) (Lower is better)



Edge ResNet-50 Multi-stream (8) Latencies (ms) (Lower is better)



Lowest Latencies - Edge Servers – By Accelerators

Lowest Latencies -> least response time at Edge

Source: MLPerf ™ 2.1 Submission ID: Alibaba 2.1-0004, 2.1-0003, Qualcomm 2.1-0107, Nvidia, H3C 2.1-0048 , 2.1-0046, 2.1-0044, HPE 2.1-0054 Supermicro 2.1-0112, 2.1-0118, Fujitsu 2.1-0020, Inspur 2.1-0058, Nettrix 2.1-0098

Edge – RetinaNet Object detector

Submission Description				RetinaNet				
ID	Submitter	System	Accelerator	SingleStream Latency (ms)	MultiStream Latency (ms)	Offline (QPS)	Power (Watts)	QPS/Watt
2.1-1564	Dell	Dell PowerEdge R7515	4x PCIe Standard	24.23	60.75	660.09	334.75	1.97
2.1-1565	HPE	HPE ProLiant e920d	4x PCIe Standard	25.39	63.67	660.94	296.49	2.23
2.1-1566	Lenovo	Lenovo ThinkSystem SE350 Edge Server	1x PCIe Pro	23.94	127.92	187.00	132.64	1.41
2.1-1575	Qualcomm	Thundercomm TurboX EB6	1x DM.2e	38.49	331.39	62.95	20.89	3.01
2.1-1576	Qualcomm	Foxconn Gloria (Highend)	1x DM.2	43.76	310.15	93.76	22.48	4.17
2.1-1577	Qualcomm	Foxconn Gloria (Entry)	1x DM.2e	38.61	326.60	67.04	21.43	3.13
2.1-1578	Qualcomm	Inventec Heimdall	1x DM.2e	39.27	332.17	63.07	19.51	3.23
2.1-1579	Qualcomm	Gigabyte R282-Z93	2x PCIe Pro	24.40	65.22	374.41	-	-
2.1-1580	Qualcomm	Gigabyte R282-Z93	5x PCIe Pro	24.31	58.14	932.89	-	-
2.1-1581	Qualcomm	Gigabyte R282-Z93	5x PCIe Pro EE*	24.29	57.47	932.07	406.92	2.29

Qualcomm Cloud AI 100 with MLPerf ™ 2.1 - RetinaNet

Raising the bar – AI Edge Power efficiency

All RetinaNet submissions are made in Open Division with Preview mode. For source refer to Qualcomm submissions in Open Edge (,power) Division
QPS – Queries/Sec = Samples/Sec. Power efficiency is not official MLPerf™ metric. It is derived from Open Edge Power submissions.

Datacenter – Open, Preview
Performance – Server, Offline
(Inference/second)

Qualcomm Cloud AI 100 PCIe HHHL/Pro			RetinaNet		
ID	Submitter	System	Server	Offline	Per Accelerator
2.1-1572	Qualcomm	Gigabyte G292 Z43 (18x Qualcomm Cloud AI 100 PCIe pro)	2,296	3,358	187
2.1-1573	Qualcomm	Gigabyte G292 Z43 (16x Qualcomm Cloud AI 100 PCIe Pro)	2046	3002	188
2.1-1573	Qualcomm	Gigabyte R282 Z93 (8x Qualcomm Cloud AI 100 PCIe Pro)	997	1,501	188
2.1-1563	Dell	PowerEdge 7515 (4x Qualcomm Cloud AI 100 PCIe Pro)	473	747	187

Datacenter – Open, Preview
Power – Offline
(Inference/second)

Qualcomm Cloud AI 100 PCIe HHHL/Pro			RetinaNet		
ID	Submitter	System	Sample/s	Watts	QPS/Watt
2.1-1574	Qualcomm	Gigabyte R282 Z93 (8x Qualcomm Cloud AI 100)	1,501	566	2
2.1-1572	Krai	Gigabyte G292 Z43 (18x Qualcomm Cloud AI 100, EE*)	3358	1777	1
2.1-1563	Dell	PowerEdge 7515 (4x Qualcomm Cloud AI 100)	747	353	2

Qualcomm Cloud AI 100 with MLPerf ™ 2.1 - RetinaNet

Performant, power efficient datacenter object detection solutions

All RetinaNet submissions are made in open division with Preview mode. For source refer to Qualcomm submissions in Open Edge (,power) Division. Power efficiency is not official MLPerf™ metric. It is derived from Open Edge Power submissions.

Thank you

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog