Qualcomm

Qualcomm Cloud AI 100 MLPerf™ Inference Benchmarks

With industry leading advancements in performance density and performance-per-watt capabilities, the Qualcomm Cloud AI 100 platforms are leading in all scorecards of the latest benchmark submissions.

Updated **September 2022**

Cloud AI 100 MLPerf™ v2.1 Inference Datacenter Performance Benchmarks

Closed Division — Cloud AI SDK Version 1.7.1

Network Type	Network	Precision	Server			Offline			Accelerator	Server	Dataset	Target Accuracy	Accuracy Achieved	MLPerf™ Submission ID	Submitter
			Queries/ sec	System Power (W)	Queries/ sec/Watt	Samples/ sec	System Power (W)	Samples/ sec/Watt							
Image Classification	ResNet-50	INT8	330,079.00			428,786.00			18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	ImageNet	75.99%	99.0%	2.1-0100	Qualcomm
			250,060.00	1,441.88	173.43	357,387.00	1,723.60	207.35						2.1-0063	Krai
			173,044.00			182,343.00			8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93				2.1-0101	Qualcomm
			151,029.00	723.19	208.84	169,680.00	781.94	217.00						2.1-0102	
			91,206.60	512.29		92,235.40	502.15		4x Cloud AI 100 PCIe/HHHL Pro	Dell PowerEdge R7515				2.1-0016	Dell
Natural Language Processing	BERT-99 BERT Large SeqLen 384	mixed	12,644.90			13,541.10			18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	SQuAD v1.1		99.0%	2.1-0100	Qualcomm
			11,595.70			12,373.30			16x Cloud AI 100 PCIe/HHHL Pro			90.17% f1		2.1-0099	
			5,794.93			6,149.99			8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93				2.1-0101	
			5,296.66	595.90	8.89	5,597.99	613.32	9.13						2.1-0102	
			2,550.92	455.25	5.60	2,913.00	476.31	6.12	4x Cloud AI 100 PCIe/HHHL Pro	Dell PowerEdge R7515				2.1-0016	Dell
Natural Language Processing	BERT-99.9 BERT Large SeqLen 384	FP16	5,894.90			6,556.65			18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	SQuAD v1.1	90.79% f1	99.9%	2.1-0100	Qualcomm
			5,494.91			5,859.43			16x Cloud AI 100 PCIe/HHHL Pro					2.1-0099	
			2,646.25			2,799.24			8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93				2.1-0101	
			2,096.39	549.22	3.82	2,445.08	590.57	4.14						2.1-0102	
			1,277.56	434.74	2.94	1,394.35	433.96	3.21	4x Cloud AI 100 PCIe/HHHL Pro	Dell PowerEdge R7515				2.1-0016	Dell

Open Division — Cloud AI SDK Version 1.8.0.73

Network Type	Network	Precision	Server			Offline			Accelerator	Server	Dataset	Target Accuracy	Accuracy Achieved	MLPerf™ Submission ID	Submitter
			Queries/ sec	System Power (W)	Queries/ sec/Watt	Samples/ sec	System Power (W)	Samples/ sec/Watt							
Object detection	(RetinaNet*)	INT8	2,296.04			3,358.26			18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	OpenImages	NA	37.161	2.1-1568	Qualcomm
			2,046.17			3,002.87			16x Cloud AI 100 PCIe/HHHL Pro					2.1-1567	
			997.75			1,501.28			8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93				2.1-1569	
			979.57	477.18	2.05	1,501.01	566.42	2.65		Gigabyte R282-Z94				2.1-1570	
			473.08	315.15	1.50	747.50	353.56	2.11	4x Cloud AI 100 PCIe/HHHL Pro	Dell PowerEdge R7515				2.1-1559	Dell

* RetinaNet Submitted to Open Division

Cloud AI 100 MLPerf™ v2.1 Inference Edge Performance Benchmarks

Closed Division Cloud AI SDK Version 1.7.1

Task	Network	Precision	Single Stream		Multi Stream (8 Streams)		Offline			Accelerator	Platform	Dataset	Target Accuracy	Accuracy Achieved	MLPerf™ Submission ID	Submitter
			Latency (ms)	System Energy (J)	Latency (ms)	System Energy (J)	Samples/ sec	System Power (W)	Samples/ sec/ Watt							
Image classification	ResNet-50	INT8	0.40		0.59		118,819.00			5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	ImageNet	76.00%	99.0%	2.1-0107	Qualcomm
			0.58	0.11	0.62	0.16	106,534.00	547.51	194.58						2.1-0108	
			0.41		0.63		47,244.00			2x Cloud AI 100 PCIe/HHHL Pro					2.1-0106	
			0.40	0.04	1.02	0.12	23,311.00	178.97	130.25	1x Cloud AI 100 PCIe/HHHL Pro	Lenovo ThinkSystem SE350 Edge Server				2.1-0081	Lenovo
			0.41	0.09	0.87	0.21	79,515.20	472.00	168.46	4x Cloud AI 100 PCIe/HHHL Standard	Dell PowerEdge R7515				2.1-0017	Dell
			0.62	0.13	0.66	0.18	77,652.40	450.99	172.18		HPE ProLiant e920d				2.1-0054	HPE
			0.86	0.01	2.51	0.05	9,224.17	29.67	310.84	1x Cloud AI 100 DM.2 20W	Foxconn Gloria (Highend)				2.1-0105	Qualcomm
			0.69	0.01	2.22	0.04	7,293.49	25.45	286.62	1x Cloud AI 100 DM.2e 15W	Foxconn Gloria (Entry)				2.1-0104	
			0.65	0.01	2.27	0.05	7,259.95	25.32	286.70		Thundercomm TurboX EB6 Edge AI Box				2.1-0103	
			0.63	0.01	2.44	0.04	6,875.00	23.12	297.30		Inventec Heimdall				2.1-0123	
Natural Language Processing	BERT-99	mixed	7.34				3,770.90			5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	SQuAD v1.1	90.17% f1	99.0%	2.1-0107	Qualcomm
			9.07	2.03			3,549.11	466.61	7.61						2.1-0108	
			7.34				1,505.22			2x Cloud AI 100 PCIe/HHHL Pro					2.1-0106	
			7.28	0.91			750.60	172.71	4.35	1x Cloud AI 100 PCIe/HHHL Pro	Lenovo ThinkSystem SE350 Edge Server				2.1-0081	Lenovo
			7.30	1.74			2,923.32	475.74	6.14	4x Cloud AI 100 PCIe/HHHL Standard	Dell PowerEdge R7515				2.1-0017	Dell
			7.31	1.76			2,863.63	418.39	6.84		HPE ProLiant e920d				2.1-0054	HPE
			11.80	0.25			369.48	31.43	11.76	1x Cloud AI 100 DM.2 20W	Foxconn Gloria (Highend)				2.1-0105	Qualcomm
			12.43	0.29			273.42	25.50	10.72	1x Cloud AI 100 DM.2e 15W	Foxconn Gloria (Entry)				2.1-0104	
			12.23	0.29			264.13	24.33	10.86		Thundercomm TurboX EB6 Edge AI Box				2.1-0103	
			13.00	0.27			254.21	22.03	11.54		Inventec Heimdall				2.1-0123	

Continued on next page

Cloud AI 100 MLPerf™ v2.1 Inference Edge Performance Benchmarks

Open Division Cloud AI SDK Version 1.8.0.73

Task	Network	Precision	Single Stream		Multi Stream (8 Streams)		Offline			Accelerator	Platform	Dataset	Target Accuracy	Accuracy Achieved	MLPerf™ Submission ID	Submitter
			Latency (ms)	System Energy (J)	Latency (ms)	System Energy (J)	Samples/ sec	System Power (W)	Samples/ sec/ Watt							
Object detection	(RetinaNet*)	INT8	24.31		58.14		932.89			5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	OpenImages	NA	37.161	2.1-1576	Qualcomm
			24.29	5.18	57.47	14.44	932.07	406.92	2.29						2.1-1577	
			24.40		65.22		374.41			2x Cloud AI 100 PCIe/HHHL Pro					2.1-1575	
			23.94	2.46	127.92	13.51	187.00	132.64	1.41	1x Cloud AI 100 PCIe/HHHL Pro	Lenovo ThinkSystem SE350 Edge Server				2.1-1562	Lenovo
			24.23	5.29	60.75	14.77	660.09	334.75	1.97	4x Cloud AI 100 PCIe/HHHL Standard	Dell PowerEdge R7515				2.1-1560	Dell
			25.39	5.46	63.67	15.47	660.94	296.49	2.23		HPE ProLiant e920d				2.1-1561	HPE
			43.76	0.61	310.15	4.58	93.76	22.48	4.17	1x Cloud AI 100 DM.2 20W	Foxconn Gloria (Highend)				2.1-1572	Qualcomm
			38.61	0.62	326.60	5.02	67.04	21.43	3.13	1x Cloud AI 100 DM.2e 15W	Foxconn Gloria (Entry)				2.1-1573	
			38.49	0.62	331.39	5.07	62.95	20.89	3.01		Thundercomm TurboX EB6 Edge AI Box				2.1-1571	
			39.27	0.59	332.17	4.63	63.07	19.51	3.23		Inventec Heimdall				2.1-1574	

* RetinaNet Submitted to Open Division

Cloud AI 100 MLPerf™ v2.1 Inference Datacenter RetinaNet* Performance Benchmarks

Cloud AI SDK Version 1.8.0.73

Task	Network	Precision	Server			Offline			Accelerator	Server	Dataset	Target Accuracy	Accuracy Achieved	MLPerf™ Submission ID	Submitter
			Queries/sec	System Power (W)	Queries/sec/ Watt	Samples/sec	System Power (W)	Samples/ sec/Watt							
Object detection	(RetinaNet*)	INT8	2,228.56			3,213.66			18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	OpenImages	mAP=37.1745	37.239	NA	Qualcomm
			2,210.23	1,296.97	1.70	3,152.26	1,486.98	2.12							
			1,979.89			2,854.05			16x Cloud AI 100 PCIe/HHHL Pro						
			1,692.84	1,168.49	1.45	2,853.74	1,405.10	2.03							
			965.55			1,436.15			8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93					
			949.67	474.52	2.00	1,436.15	567.62	2.53							
			455.11	320.78	1.42	717.36	360.97	1.99	4x Cloud AI 100 PCIe/HHHL Pro	Dell PowerEdge R7515					Dell

* RetinaNet Benchmark Results Based on MLPerf™ v2.1 branch - Updated with improved Accuracy bettering Closed Div requirements

Cloud AI 100 MLPerf™ v2.1 Inference Edge Performance RetinaNet* Benchmarks

Cloud AI SDK Version 1.8.0.73

Task	Network	Precision	Single Stream		Multi Stream (8 Streams)		Offline			Accelerator	Platform	Dataset	Target Accuracy	Accuracy Achieved	MLPerf™ Submission ID	Submitter
			Latency (ms)	System Energy (J)	Latency (ms)	System Energy (J)	Samples/ sec	System Power (W)	Samples/ sec/Watt							
Object detection	(RetinaNet*)	INT8	24.93		57.85		896.53			5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	OpenImages	mAP=37.1745	37.239	NA	Qualcomm
			24.92	5.32	58.14	14.86	896.95	418.80	2.14							
			24.64		63.31		358.27			2x Cloud AI 100 PCIe/HHHL Pro						
			24.51	2.54	123.03	13.36	179.36	136.07	1.32	1x Cloud AI 100 PCIe/HHHL Pro	Lenovo ThinkSystem SE350					Lenovo
			24.80	5.37	56.84	14.19	643.23	344.60	1.87	4x Cloud AI 100 PCIe/HHHL Standard	Dell PowerEdge R7515					Dell
			25.98	5.52	58.15	14.43	643.36	300.29	2.14		HPE ProLiant e920d					HPE
			44.18	0.61	311.62	4.62	90.32	22.29	4.05	1x Cloud AI 100 DM.2 20W	Foxconn Gloria (Highend)			37.190		Qualcomm
			36.90	0.57	324.11	5.03	67.67	21.63	3.13	1x Cloud AI 100 DM.2e 15W	Foxconn Gloria (Entry)					
			37.38	0.57	329.18	5.04	65.07	21.52	3.02		Thundercomm TurboX EB6 Edge AI Box					
			37.31	0.54	332.05	4.66	64.03	19.97	3.21		Inventec Heimdall					

* RetinaNet Benchmark Results Based on MLPerf™ v2.1 branch - Updated with improved Accuracy bettering Closed Div requirements