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 **May 2022**

Qualcomm Cloud AI 100 MLPerf™ v2.0 Inference Datacenter Performance Benchmarks

Network Category	Network	Precision	Accuracy (%FP32 ref)	Server			Offline			MLPerf™ Submission ID	Submitter	Accelerator	Server	Cloud AI SDK	Target Accuracy	Dataset
				queries/sec	System Power (W)	queries/sec/Watt	samples/sec	System Power (W)	*Power Efficiency (Inf/S/Wat)							
Image classification	ResNet-50	INT8	99.0%	351057.00			418794.00			2.0-123	Qualcomm	18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80	75.99%	ImageNet
				273027.00	1404.18	194.44	337433.00	1605.89	210.12	2.0-124						
				327119.00			371473.00			2.0-034	GIGABYTE	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80		
				265029.00	1348.78	196.50	311873.00	1532.39	203.52	2.0-035						
				180003.00			181477.00			2.0-125	Qualcomm	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				150999.00	663.07	227.73	155761.00	677.23	230.00	2.0-126						
				176002.00	828.74	212.37	179188.00	823.93	217.48	2.0-054	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93			
				112999.00			114150.00			2.0-052	KRAI	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				94996.90	484.51	196.07	97527.30	491.46	198.44	2.0-053						
Object detection (large)	SSD-Large SSD-Resnet-34	INT8	99.0%	8044.24			8127.37			2.0-123	Qualcomm	18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80	19.827% mAP	COCO
				6098.20	1581.31	3.86	6822.00	1667.38	4.09	2.0-124						
				7185.02			7236.82			2.0-034	GIGABYTE	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80		
				6078.03	1493.62	4.07	6155.78	1507.10	4.08	2.0-035						
				3567.17			3611.55			2.0-125	Qualcomm	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				2997.58	629.19	4.76	3110.26	640.99	4.85	2.0-126						
				3527.25	827.00	4.27	3607.83	830.84	4.34	2.0-054	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				2147.98			2223.10			2.0-052	KRAI	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				1849.43	466.62	3.96	1927.75	479.15	4.02	2.0-053						
Natural Language Processing	BERT-99 BERT Large SeqLen 384	mixed	99.0%	12738.90			13223.90			2.0-123	Qualcomm	18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80	90.17% f1	SQuAD v1.1
				10793.50	1486.39	7.26	11305.60	1439.02	7.86	2.0-124						
				11313.90			11777.90			2.0-034	GIGABYTE	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80		
				9594.98	1361.90	7.05	10045.80	1371.41	7.33	2.0-035						
				5546.79			5796.53			2.0-125	Qualcomm	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				4697.46	575.00	8.17	5049.70	593.68	8.51	2.0-126						
				5511.75	790.62	6.97	5745.17	794.30	7.23	2.0-054	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				3297.86			3597.91			2.0-052	KRAI	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				2848.10	424.31	6.71	3142.52	438.99	7.16	2.0-053						
Natural Language Processing	BERT-99.9 BERT Large SeqLen 384	FP16	99.9%	6297.67			6560.78			2.0-123	Qualcomm	18x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80	90.79% f1	SQuAD v1.1
				4797.61	1309.15	3.66	5254.36	1294.43	4.06	2.0-124						
				5548.77			5824.97			2.0-034	GIGABYTE	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.6.80		
				4448.17	1215.17	3.66	4629.87	1203.72	3.85	2.0-035						
				2697.85			2906.18			2.0-125	Qualcomm	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				2098.48	525.55	3.99	2335.47	547.81	4.26	2.0-126						
				2667.85	723.84	3.69	2906.84	740.28	3.93	2.0-054	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				1619.50			1769.71			2.0-052	KRAI	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				1300.33	397.71	3.27	1448.92	407.98	3.55	2.0-053						

* Power efficiency at total System power

Qualcomm Cloud AI 100 MLPerf™ v2.0 Inference Edge Performance Benchmarks

Network Category	Network	Precision	Accuracy (%FP32 ref)	Single Stream		Multi Stream (8 streams)		Offline			MLPerf™ Submission ID	Submitter	Accelerator	Platform	Cloud AI SDK	Target Accuracy	Dataset
				latency (ms)	System energy (J) per stream	latency (ms)	System energy (J) per stream	samples/ sec	System Power (W)	*Power Efficiency (Inf/S/Wat)							
Image classification	ResNet-50	INT8	99.0%	0.89	0.02	3.30	0.06	9780.43	35.95	272.02	2.0-128	Qualcomm	1x Cloud AI 100 DM.2 20W TDP	AI Edge Development Kit	v1.6.80	76.0%	ImageNet
				0.84	0.02	3.91	0.06	5849.30	24.00	243.77	2.0-127	Qualcomm	1x Cloud AI 100 DM.2e 15W TDP	AI Edge Development Kit	v1.6.80		
				0.75		2.52		7047.83			2.0-003	Alibaba	1x Cloud AI 100 DM.2e 15W TDP	Alibaba Cloud Sinian Platform (Haishen)	v1.6.80		
				0.82	0.01	3.86	0.05	7309.24	24.34	300.32	2.0-129	Qualcomm	1x Cloud AI 100 DM.2e 15W TDP	Foxconn Gloria	v1.6.80		
				0.35		1.02		23807.50			2.0-130	Qualcomm	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.34		0.64		46361.40			2.0-036	GIGABYTE	2x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.34		0.42		114869.00			2.0-131	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.38	0.08	0.46	0.12	97426.10	488.03	199.63	2.0-132	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z94	v1.6.81		
				0.33		0.42		179618.00			2.0-059	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.34	0.09	0.47	0.15	155740.00	665.34	234.08	2.0-060	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z94	v1.6.80		
Object detection (small)	SSD-small SSD-MobileNet v1	INT8	99.0%	1.74	0.03	5.87	0.10	5205.54	35.01	148.67	2.0-128	Qualcomm	1x Cloud AI 100 DM.2 20W TDP	AI Edge Development Kit	v1.6.80	23.163% mAP	COCO
				1.71	0.03	6.41	0.10	3477.95	25.26	137.68	2.0-127	Qualcomm	1x Cloud AI 100 DM.2e	AI Edge Development Kit	v1.6.80		
				1.72		8.45		3949.22			2.0-003	Alibaba	1x Cloud AI 100 DM.2e	Alibaba Cloud Sinian Platform (Haishen)	v1.6.80		
				1.99	0.03	7.29	0.10	4002.56	25.94	154.28	2.0-129	Qualcomm	1x Cloud AI 100 DM.2e 15W TDP	Foxconn Gloria	v1.6.80		
				0.69		1.85		19616.40			2.0-130	Qualcomm	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.68		1.52		38630.30			2.0-036	GIGABYTE	2x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.68		1.77		95831.50			2.0-131	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.74	0.16	1.77	0.34	82774.10	553.35	149.59	2.0-132	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.65		0.94		118589.00			2.0-059	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				0.73	0.17	0.98	0.26	116636.00	693.21	168.25	2.0-060	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		

* Power efficiency at total System power

Qualcomm Cloud AI 100 MLPerf™ v2.0 Inference Edge Performance Benchmarks (Cont.)

Network Category	Network	Precision	Accuracy (%FP32 ref)	Single Stream		Multi Stream (8 streams)		Offline			MLPerf™ Submission ID	Submitter	Accelerator	Platform	Cloud AI SDK	Target Accuracy	Dataset
				latency (ms)	System energy (J) per stream	latency (ms)	System energy (J) per stream	samples/ sec	System Power (W)	*Power Efficiency (Inf/S/Wat)							
Object detection (large)	SSD-Large SSD-ResNet-34	INT8	99.0%	22.21	0.42	96.91	2.50	213.59	38.65	5.53	2.0-128	Qualcomm	1x Cloud AI 100 DM.2 20W TDP	AI Edge Development Kit	v1.6.80	19.827% mAP	COCO
				21.54	0.42	91.86	2.41	224.65	42.84	5.24	2.0-055	KRAI	1x Cloud AI 100 DM.2 25W TDP	AI Edge Development Kit	v1.6.80		
				25.68	0.44	129.49	2.57	128.77	26.44	4.87	2.0-127	Qualcomm	1x Cloud AI 100 DM.2e	AI Edge Development Kit	v1.6.80		
				26.97		120.24		150.60			2.0-003	Alibaba	1x Cloud AI 100 DM.2e	Alibaba Cloud Sinian Platform (Haishen)	v1.6.80		
				27.07	0.40	114.14	2.25	163.42	28.44	5.75	2.0-129	Qualcomm	1x Cloud AI 100 DM.2e 15W TDP	Foxconn Gloria	v1.6.80		
				8.94		43.45		443.16			2.0-130	Qualcomm	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				8.73		28.03		885.04			2.0-036	GIGABYTE	2x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				8.61		25.94		2224.46			2.0-131	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				9.48	1.95	28.75	6.89	1927.92	472.96	4.08	2.0-132	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				9.02		12.65		3619.98			2.0-059	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				9.43	2.12	13.28	3.94	3109.66	640.92	4.08	2.0-060	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
Natural Language Processing	BERT-99	mixed	99.0%	15.41	0.45			315.89	36.61	8.63	2.0-128	Qualcomm	1x Cloud AI 100 DM.2 20W TDP	AI Edge Development Kit	v1.6.80	90.17% f1	SQuAD v1.1
				17.38	0.45			361.09	39.75	9.09	2.0-055	KRAI	1x Cloud AI 100 DM.2 25W TDP	AI Edge Development Kit	v1.6.80		
				16.38	0.35			179.47	22.73	7.89	2.0-127	Qualcomm	1x Cloud AI 100 DM.2e	AI Edge Development Kit	v1.6.80		
				14.42				220.61			2.0-003	Alibaba	1x Cloud AI 100 DM.2e	Alibaba Cloud Sinian Platform (Haishen)	v1.6.80		
				13.73	0.29			230.30	23.72	9.71	2.0-129	Qualcomm	1x Cloud AI 100 DM.2e 15W TDP	Foxconn Gloria	v1.6.80		
				10.27				724.08			2.0-130	Qualcomm	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				10.25				1437.71			2.0-036	GIGABYTE	2x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				10.25				3606.50			2.0-131	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				10.96	2.52			3142.44	436.11	7.21	2.0-132	Qualcomm	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				10.41				5783.39			2.0-059	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		
				10.81	2.74			5049.90	591.15	8.54	2.0-060	KRAI	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.6.80		

* Power efficiency at total System power