



Qualcomm Cloud AI 100 Performance 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 **January 2022**



PCIe (HHHL), ECC Enabled: Transformer Based Models

Network	Layers	Sequence Length	Precision	Batch Size	Throughput sequences/sec	Latency (ms)	Accelerator	Server	Cloud AI SDK
DistilBERT	6	128	FP16	8	3853	8.41	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
BERT-BASE	12	128	FP16	8	1856	16.80	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
BERT-BASE	12	128	mixed	8	4083	16.23	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
BERT-LARGE	24	128	FP16	4	545	29.09	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
BERT-LARGE	24	128	mixed	8	1176	28.21	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
BERT-LARGE	24	384	FP16	1	311	26.02	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
BERT-LARGE	24	384	mixed	1	681	23.15	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
DistilRoBERTa-BASE	6	128	FP16	8	3633	8.81	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
DistilGPT2	6	32	FP16	32	11920	10.54	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
OpenAI-CLIP	12	77	FP16	2	709	22.78	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1
MiniLM	12	128	FP16	8	5004	6.76	1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.1



MLPerf Inference Datacenter v1.1 Performance Benchmarks

Task	Network	Precision	Accuracy (%FP32 ref)	Server		Offline		Accelerator	Server	Cloud AI SDK	Dataset	Target Accuracy
				queries/sec	System Power (W)	samples/sec	System Power (W)					
Image classification	ResNet	INT8	99.0%	310,064.00	1,790.98	342,011.00	1,844.48	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.6	ImageNet	75.99%
			99.0%	144,521.00	812.05	169,231.00	857.29	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
Object detection (large)	SSD-Large	INT8	99.0%	6,868.05	1,917.73	6,958.41	1,927.84	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.6	COCO	19.827% mAP
			99.0%	3,381.99	840.13	3,469.93	862.79	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
Natural Language Processing	BERT	mixed	99.0%	10,303.30	1,732.29	10,601.90	1,712.09	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.9	SQuAD v1.1	90.17% f1
			99.0%	4,902.82	765.17	5,202.88	776.30	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.9		
Natural Language Processing	BERT	FP16	99.9%	5,102.64	1,819.17	5,508.44	1,881.14	16x Cloud AI 100 PCIe/HHHL Pro	Gigabyte G292-Z43	v1.5.9	SQuAD v1.1	90.79% f1
			99.9%	2,250.17	763.92	2,664.14	841.44	8x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.9		



MLPerf Inference Edge v1.1 Performance Benchmarks

Task	Network	Precision	Accuracy (%FP32 ref)	Single Stream		Offline		Accelerator	Server	Cloud AI SDK	Dataset	Target Accuracy
				latency (ms)	System energy (J) per stream	samples/sec	System Power (W)					
Image classification	ResNet	INT8	99.0%	1.03	0.02	5,729.06	23.79	1x Cloud AI 100 DM.2e	AI Development Kit	v1.4.66	ImageNet	76.00%
			99.0%	1.02	0.02	8,910.85	37.14	1x Cloud AI 100 DM.2	AI Development Kit	v1.4.66		
			99.0%	1.62		22,667.00		1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
			99.0%	1.62	0.35	107,907.00	618.11	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
Object detection (small)	SSD-small	INT8	99.0%	1.90	0.04	3,438.13	24.26	1x Cloud AI 100 DM.2e	AI Development Kit	v1.4.66	COCO	23.163% mAP
			99.0%	2.00	0.04	4,091.27	31.80	1x Cloud AI 100 DM.2	AI Development Kit	v1.4.66		
			99.0%	3.90		19,363.80		1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
			99.0%	3.90	0.87	97,514.70	692.84	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
Object detection (large)	SSD-Large	INT8	99.0%	27.19	0.48	120.11	24.36	1x Cloud AI 100 DM.2e	AI Development Kit	v1.4.66	COCO	19.827% mAP
			99.0%	27.52	0.49	178.66	37.08	1x Cloud AI 100 DM.2	AI Development Kit	v1.4.66		
			99.0%	39.59		435.03		1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
			99.0%	39.66	9.56	2,175.14	593.49	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.6		
Natural Language Processing	BERT	mixed	99.0%	11.32		670.91		1x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.9	SQuAD v1.1	90.17% f1
			99.0%	11.36	2.78	3,279.03	552.42	5x Cloud AI 100 PCIe/HHHL Pro	Gigabyte R282-Z93	v1.5.9		

MLPerf™ v1.1 Inference Closed Edge; data derived from submissions: Qualcomm 1.1-120, 1.1-121, 1.1-122, 1.1-123, 1.1-124, 1.1-125