Qualcomm

Qualcomm Cloud AI 100 Inference Performance

With industry leading advancements in performance density and performance-per-watt capabilities, the Qualcomm Cloud AI 100 Accelerators are leading in all inference performance metrics.

Updated **May 2023**

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Pro (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Natural Language Processing

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
ALBERT-BASE	12	128	fp16	1	Latency	725	2.7124	760	17
		128	fp16	4	Throughput	2250	14.091	2196	33
		128	mixed	8	Throughput	4468	14.2502	4402	68
BERT-BASE	12	128	fp16	1	Latency	715	2.752	743	16
		128	fp16	4	Throughput	2305	13.7998	2263	34
		128	mixed	1	Latency	1197	1.6265	1354	35
		128	mixed	8	Throughput	4523	13.9206	4458	69
		256	fp16	1	Latency	354	2.76	343	10
		256	fp16	4	Throughput	1054	14.9941	1059	17
		512	fp16	1	Latency	176	5.596	181	5
		512	fp16	1	Throughput	481	16.4859	477	7
BERT-LARGE	24	128	fp16	1	Latency	169	5.8334	174	5
		128	fp16	8	Throughput	633	25.3226	618	10
		128	mixed	1	Latency	380	5.1794	393	9
		128	mixed	8	Throughput	1445	21.5869	1434	23
		256	fp16	1	Latency	135	7.3238	137	3
		256	fp16	4	Throughput	339	46.8281	339	5
		384	fp16	1	Latency	95	10.3381	97	2
		384	fp16	1	Throughput	170	23.338	174	3
		512	fp16	1	Latency	74	13.4071	75	2
		512	fp16	1	Throughput	135	29.4397	135	2
BERT-LARGE-PACKEDBOOL	24	384	fp16	1	Throughput	359	23.1446	353	6
		384	mixed	1	Balanced	646	12.8449	654	11
		384	mixed	1	Throughput	752	44.4194	742	11
		448	fp16	1	Throughput	367	52.8736	369	6
		448	mixed	1	Throughput	724	26.6545	714	11
DeBERTa3xs (comp mask)	12	2048	fp16	1	Throughput	5	178.3358	5	0
DeBERTaV3xs	12	128	fp16	1	Latency	789	2.4921	837	31
		128	fp16	4	Throughput	3660	4.2567	3796	59

Continued on next page

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Pro (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Natural Language Processing

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
DistilBERT	6	128	fp16	1	Latency	1413	1.4131	1661	44
		128	fp16	8	Throughput	5034	12.4576	4926	75
		128	mixed	1	Latency	3596	1.0234	4134	81
		128	mixed	16	Throughput	8617	7.3312	8744	151
DistilGPT2	6	32	fp16	32	Throughput	12067	10.5545	13724	250
		8	fp16	64	Throughput	28569	18.0089	33233	639
DistillRoBERTa-BASE	6	128	fp16	1	Latency	1437	1.393	1651	44
		128	fp16	8	Throughput	5077	12.5409	4905	74
GPT2-Small	6	8	fp16	1	Latency	456	2.2022	588	18
		8	fp16	16	Throughput	2421	13.2327	4442	118
MiniLM	12	128	fp16	1	Latency	2783	1.4219	3181	57
		128	fp16	8	Throughput	5746	11.0092	5768	92
OpenAI-CLIP	12	77	fp16	1	Latency	312	6.363	321	7
		77	fp16	4	Throughput	911	17.4504	917	16
ROBERTA-BASE	12	128	fp16	1	Latency	720	2.7195	743	16
		128	fp16	4	Throughput	2357	13.4729	2343	36
ROBERTA-LARGE	24	128	fp16	1	Latency	165	5.9756	171	5
		128	fp16	8	Throughput	633	25.0903	639	10

Continued on next page

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Pro (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Vision Transformer

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
VITBASE	12	224x224	fp16	1	Latency	689	1.4138	822	23
		224x224	fp16	12	Throughput	5824	16.3681	5822	87

Computer Vision - Image Classification

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
EFFICIENTNETB0		224x224	int8	1	Latency	6498	0.6065	9779	276
			int8	1	Throughput	15872	1.0151	21389	382
			fp16	1	Latency	3898	1.0169	5000	109
			fp16	1	Throughput	7203	2.2395	7849	123
RESNET101	101	224x224	int8	1	Latency	3244	0.6054	4502	106
		224x224	int8	16	Throughput	11325	5.5262	13311	208
		224x224	fp16	1	Latency	1391	1.4134	1692	33
		224x224	fp16	8	Throughput	4260	7.33	3869	59
RESNET50-V1.5	50	224x224	int8	8	Throughput	10818	2.9018	22403	339

* Power efficiency measured at accelerator level

^ Network compilation Optimized for --

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Pro (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Computer Vision - Object Detection

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
EFFICIENTDET		224x224	fp16	1	Latency	131	7.6283	221	13
			fp16	1	Throughput	561	14.112	728	18
MASK-RCNN-COCO		416x416	fp16	1	Latency	40	24.8546	47	2
		416x416	fp16	1	Throughput	140	56.5666	147	3
		416x416	mixed	1	Latency	62	15.9682	82	4
		416x416	mixed	1	Throughput	267	29.6941	310	7
RESNET34_SSD	34	1200x1200	mixed	1	Throughput	446	35.3984	486	7
RETINANET		800x800	mixed	1	Latency	52	18.843	55	2
			mixed	1	Throughput	291	55.5003	293	5
SSDMOBILENETV1		300x300	int8	4	Throughput	17851	3.4569	24517	389
YOLOV3	106	416x416	int8	1	Latency	458	2.1884	756	19
		416x416	int8	4	Throughput	1427	10.9969	1842	29
		416x416	fp16	1	Latency	286	3.3951	379	9
	106	416x416	fp16	2	Throughput	705	11.2709	763	12
YOLOV4-LR		416x416	int8	1	Latency	363	2.6839	533	17
		416x416	int8	1	Throughput	1349	11.7549	1367	21
		416x416	fp16	1	Latency	246	3.9449	299	8
		416x416	fp16	1	Throughput	628	12.62	648	11
		416x416	mixed	1	Latency	297	3.286	394	12
		416x416	mixed	1	Throughput	1037	7.5824	1193	20
YOLOV5S	213	416x416	int8	1	Latency	1195	1.6158	2185	77
		416x416	int8	1	Throughput	6014	2.6145	8479	132
		416x416	fp16	1	Latency	508	1.9111	914	30
	213	416x416	fp16	1	Throughput	3086	5.0268	4357	71

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Standard (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Natural Language Processing

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
ALBERT-BASE	12	128	fp16	1	Latency	472	2.0425	511	15
			fp16	4	Throughput	2083	13.2702	2122	34
			mixed	8	Throughput	4069	13.6047	4039	66
BERT-BASE	12	128	fp16	1	Latency	470	2.0421	522	15
		128	fp16	4	Throughput	2161	12.8125	2165	35
		128	mixed	1	Latency	1094	1.8211	1244	36
		128	mixed	8	Throughput	4030	13.8737	4081	67
		256	fp16	1	Latency	371	2.637	393	9
		256	fp16	2	Throughput	956	14.5218	967	16
		512	fp16	1	Latency	196	5.0295	200	5
		512	fp16	1	Throughput	436	15.9304	440	7
BERT-LARGE	24	128	fp16	1	Latency	154	6.3994	158	4
		128	fp16	4	Throughput	549	21.8475	552	9
		128	mixed	1	Latency	230	4.2713	243	9
		128	mixed	8	Throughput	1227	19.4064	1237	21
		256	fp16	1	Latency	110	8.9759	112	3
		256	fp16	4	Throughput	286	41.3557	289	5
		384	fp16	1	Latency	83	11.9664	84	2
		384	fp16	1	Throughput	149	19.5523	150	3
		512	fp16	1	Latency	67	14.638	69	2
		512	fp16	1	Throughput	112	26.3859	113	2
BERT-LARGE-PACKEDBOOL	24	384	fp16	1	Throughput	309.32	19.8107	313.5	5
		384	mixed	1	Balanced	553.85	11.1695	566.39	10
		384	mixed	1	Throughput	700.15	41.5586	698.06	11
		448	fp16	1	Throughput	340.2	49.6335	342.63	6
		448	mixed	1	Throughput	646.38	26.2787	656.1	11
DeBERTa3xs (comp mask)	12	2048	fp16	1	Throughput	5	178.9804	5	0
DeBERTaV3xs	12	128	fp16	1	Latency	782	2.4896	833	32
		128	fp16	4	Throughput	3618	7.659	3715	62

Continued on next page

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Standard (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Natural Language Processing

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
DistilBERT	6	128	fp16	1	Latency	1461	1.3829	1652	35
		128	fp16	8	Throughput	4566	12.1937	4515	74
		128	mixed	1	Latency	2824	1.0202	3234	86
		128	mixed	16	Throughput	7217	6.5623	7439	133
DistilGPT2	6	32	fp16	32	Throughput	9658	10.0339	11012	216
		8	fp16	64	Throughput	26174	17.1545	31594	656
DistillRoBERTa-BASE	6	128	fp16	1	Latency	1479	1.3611	1697	38
		128	fp16	8	Throughput	4505	12.357	4589	75
GPT2-Small	6	8	fp16	1	Latency	395	2.4453	513	18
		8	fp16	14	Throughput	2316	11.8579	4316	126
MiniLM	12	128	fp16	1	Latency	2108	1.4205	2453	58
		128	fp16	8	Throughput	5013	11.0614	5100	86
OpenAI-CLIP	12	77	fp16	1	Latency	317	6.2435	328	7
		77	fp16	4	Throughput	746	15.9292	766	14
ROBERTA-BASE	12	128	fp16	1	Latency	457	2.2019	510	15
		128	fp16	8	Throughput	1900	12.8347	1940	33
	24	128	fp16	1	Latency	149	6.6201	154	4
		128	fp16	4	Throughput	552	21.768	554	9

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Standard (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Vision Transformer

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
VITBASE	12	224x224	fp16	1	Latency	690	1.4246	808	24
		224x224	fp16	12	Throughput	5333	15.6384	5433	88

Computer Vision - Image Classification

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
EFFICIENTNETB0		224x224	int8	1	Latency	4870	0.6062	7331	281
			int8	1	Throughput	14030	1.0185	18812	410
			fp16	1	Latency	2922	1.0175	3759	110
			fp16	1	Throughput	6388	2.2175	7190	122
RESNET101	101	224x224	int8	1	Latency	2441	0.8113	3699	108
		224x224	int8	16	Throughput	9079	5.2114	11059	181
		224x224	fp16	1	Latency	1119	1.8035	1289	29
		224x224	fp16	8	Throughput	3526	6.6893	3833	63
RESNET50-V1.5	50	224x224	int8	8	Throughput	8225	2.8092	18759	308

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Standard (75W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Computer Vision - Object Detection

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
EFFICIENTDET		224x224	fp16	1	Latency	137	7.4758	225	14
		416x416	fp16	1	Throughput	514	13.4399	681	18
MASK-RCNN-COCO		416x416	fp16	1	Latency	40	24.8115	47	2
		416x416	fp16	1	Throughput	130	53.3901	137	3
		416x416	mixed	1	Latency	63	15.3863	81	4
		416x416	mixed	1	Throughput	235	29.5001	274	7
RESNET34_SSD	34	1200x1200	mixed	1	Throughput	399	34.594	442	7
RETINANET		800x800	mixed	1	Latency	52	18.8625	56	3
			mixed	1	Throughput	282	49.5191	285	6
SSDMOBILENETV1		300x300	int8	4	Throughput	15959	3.4032	21949	378
YOLOV3	106	416x416	int8	1	Latency	463	2.0513	731	20
		416x416	int8	4	Throughput	1164	10.1129	1547	26
		416x416	fp16	1	Latency	268	3.6665	350	7
		416x416	fp16	1	Throughput	624	11.0423	658	11
YOLOV4-LR		416x416	int8	1	Latency	354	2.6401	502	16
		416x416	int8	1	Throughput	1227	11.3142	1281	21
		416x416	fp16	1	Latency	237	4.1096	294	8
		416x416	fp16	1	Throughput	592	11.6746	622	11
		416x416	mixed	1	Latency	312	3.0685	413	11
		416x416	mixed	1	Throughput	913	7.5097	1080	19
YOLOV5S	213	416x416	int8	1	Latency	1150	1.7158	2189	80
		416x416	int8	1	Throughput	5303	2.5929	7487	134
		416x416	fp16	1	Latency	519	1.8498	921	31
		416x416	fp16	1	Throughput	2949	4.5626	4089	71

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Lite (35 - 55W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Natural Language Processing

* All measurements at default 55W TDP settings

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
ALBERT-BASE	12	128	fp16	1	Latency	274	3.5301	294	14
		128	fp16	8	Throughput	1074	14.793	1098	39
		128	mixed	8	Throughput	1726	18.4383	1747	68
BERT-BASE	12	128	fp16	1	Latency	299	3.2785	315	14
		128	fp16	9	Throughput	1237	21.7055	1256	38
		128	mixed	1	Latency	439	2.2336	475	26
		128	mixed	8	Throughput	1780	17.8851	1805	71
		256	fp16	1	Latency	200	4.9262	208	9
		256	fp16	4	Throughput	467	16.9827	473	18
		512	fp16	1	Latency	122	8.0908	125	5
		512	fp16	1	Throughput	205	19.087	207	8
BERT-LARGE	24	128	fp16	1	Latency	87	11.3333	89	4
		128	fp16	9	Throughput	378	23.7141	382	12
		128	mixed	1	Latency	146	6.7636	150	8
		128	mixed	8	Throughput	586	27.1978	592	23
		256	fp16	1	Latency	75	13.2675	76	3
		256	fp16	4	Throughput	153	52.063	154	6
		384	fp16	1	Latency	51	19.1847	52	2
		384	fp16	1	Throughput	73	26.0633	74	3
		512	fp16	1	Latency	47	20.8496	48	2
		512	fp16	1	Throughput	58	34.0899	59	2
BERT-LARGE-PACKEDBOOL	24	384	fp16	1	Throughput	152.57	26.1413	154.66	6
		384	mixed	1	Balanced	261.25	15.8433	263.34	11
		384	mixed	1	Throughput	332.31	48.8387	336.49	14
		448	fp16	1	Throughput	162.81	56.354	162.81	6
		448	mixed	1	Throughput	281.88	33.8356	286.74	11
DeBERTa3xs	12	2048	fp16	1	Throughput	2	350.506	2	0
DeBERTaV3xs	12	128	fp16	1	Latency	741	2.6318	788	38
		128	fp16	4	Throughput	1490	5.2581	1536	60

Continued on next page

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Lite (35 - 55W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Natural Language Processing

* All measurements at default 55W TDP settings

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
DistilBERT	6	128	fp16	1	Latency	595	1.6225	671	32
		128	fp16	8	Throughput	2280	13.9463	2314	85
		128	mixed	1	Latency	1361	1.4223	1544	74
		128	mixed	16	Throughput	3446	9.2073	3522	152
DistilGPT2	6	32	fp16	32	Throughput	5083	12.4776	5998	224
		8	fp16	64	Throughput	13261	19.1805	15422	603
DistillRoBERTa-BASE	6	128	fp16	1	Latency	605	1.6238	681	31
		128	fp16	8	Throughput	2300	13.8134	2329	82
GPT2-Small	6	8	fp16	1	Latency	214	4.601	239	14
		8	fp16	16	Throughput	1244	12.8242	1932	111
MiniLM	12	128	fp16	1	Latency	1008	2.0052	1137	54
		128	fp16	8	Throughput	2205	14.456	2233	87
OpenAI-CLIP	12	77	fp16	1	Latency	124	7.9434	129	6
		77	fp16	4	Throughput	397	20.0256	404	16
ROBERTA-BASE	12	128	fp16	1	Latency	298	3.3014	312	15
		128	fp16	8	Throughput	1112	14.2757	1129	41
	24	128	fp16	1	Latency	87	11.3528	89	5
		128	fp16	9	Throughput	378	23.6953	381	12

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Lite (35 - 55W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Vision Transformer

* All measurements at default 55W TDP settings

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
VITBASE	12	224x224	fp16	1	Latency	317	3.0865	343	17
		224x224	fp16	12	Throughput	2549	18.5953	2656	98

Computer Vision - Image Classification

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
EFFICIENTNETB0		224x224	int8	1	Latency	3218	0.6084	4282	205
			int8	1	Throughput	7794	1.0171	9530	356
			fp16	1	Latency	1319	1.4546	1574	90
			fp16	1	Throughput	2685	2.9304	2828	122
RESNET101	101	224x224	int8	1	Latency	1923	1.018	2398	96
		224x224	int8	16	Throughput	5642	5.6017	7204	197
		224x224	fp16	1	Latency	538	1.8214	637	29
		224x224	fp16	8	Throughput	2050	7.7132	2233	73
RESNET50-V1.5	50	224x224	int8	8	Throughput	4998	3.1342	11729	343

* Power efficiency measured at accelerator level

^ Compile time optimization

Cloud AI 100 Inference Benchmarks - 1x PCIE HHHL Lite (35 - 55W TDP) Accelerators

Platform: Gigabyte G292-Z43 Cloud SDK Version 1.9

Computer Vision - Object Detection * All measurements at default 55W TDP settings

Network	Layers	Sequence Length	Precision	Batch Size	^Optimized for	Optimized Latency Performance		Peak Performance with Power	
						Throughput (inf/s)	E2E Latency (ms)	Throughput (inf/s)	*Power Efficiency (inf/S/W)
EFFICIENTDET		224x224	fp16	1	Latency	109	9.2032	158	8
		224x224	fp16	1	Throughput	240	16.3654	304	15
MASK-RCNN-COCO		416x416	fp16	1	Latency	30	33.0399	34	2
		416x416	fp16	1	Throughput	58	67.8535	61	3
		416x416	mixed	1	Latency	44	22.5523	53	3
		416x416	mixed	1	Throughput	102	39.0079	114	7
RESNET34_SSD	34	1200x1200	mixed	1	Throughput	188	47.4577	203	8
RETINANET		800x800	mixed	1	Latency	34	29.2458	35	2
			mixed	1	Throughput	128	69.9086	129	6
SSDMOBILENETV1		300x300	int8	4	Throughput	9372	3.7427	12463	406
YOLOV3	106	416x416	int8	1	Latency	339	2.8359	501	19
		416x416	int8	4	Throughput	749	10.5734	1014	28
		416x416	fp16	1	Latency	190	5.2328	224	9
		416x416	fp16	1	Throughput	284	6.8445	314	10
YOLOV4-LR		416x416	int8	1	Latency	273	3.4882	352	15
		416x416	int8	1	Throughput	657	11.8044	693	23
		416x416	fp16	1	Latency	172	5.7301	202	10
		416x416	fp16	1	Throughput	290	13.5062	303	11
		416x416	mixed	1	Latency	206	4.7743	250	13
		416x416	mixed	1	Throughput	402	9.8482	449	20
YOLOV5S	213	416x416	int8	1	Latency	560	1.8093	1138	58
		416x416	int8	1	Throughput	2845	2.7084	3794	134
		416x416	fp16	1	Latency	399	2.4522	612	35
		416x416	fp16	1	Throughput	1487	5.2672	1916	77

Continued on next page

* Power efficiency measured at accelerator level

^ Compile time optimization