_	BF16 CC-off vllm					BF16 CC-on vllm				
1 -	0.8	0.8	0.8	8.0	4	0.7	0.7	0.7	0.7	
2 -	3.2	3.2	3.1	3.0	4	2.8	2.9	2.8	2.7	
4 -	12.5	12.3	12.0	11.1	4	11.1	11.1	10.8	10.1	
8 -	40.8	39.8	37.2	33.6	4	42.8	41.7	39.0	34.7	
16 -	152.4	143.6	129.7	106.9	4	139.4	131.5	119.4	99.3	
32 -	524.7	474.0	398.1	301.9	4	482.8	440.9	375.4	287.2	
64 -	1628.6	1405.6	1096.3	748.7	4	1524.7	1324.0	1043.9	725.5	
128 -	4619.3	3715.6	2593.1	1666.7	4	4158.1	3416.0	2538.8	1646.4	
Sequence Length AWQ CC-off vllm					Sequence Length AWQ CC-on vllm					
1 -	1.1	1.1	1.0	1.0	4	0.9	0.9	0.9	0.9	
2 -	4.3	4.2	4.0	3.7	-	3.7	3.6	3.5	3.2	
4 -	16.5	15.8	14.6	12.6	-	14.4	13.7	12.9	11.3	
8 -	60.8	56.2	48.8	38.1	4	52.6	49.4	43.8	35.1	
16 -	208.4	182.3	145.8	104.8	4	185.1	163.7	134.0	97.9	
32 -	587.2	487.0	367.4	243.3	4	535.6	451.1	346.6	232.2	
64 -	1503.3	1198.2	847.6	532.7	4	1425.7	1137.9	808.9	514.2	
128 -	3721.9	2809.7	1889.8	1127.6	4	3578.4	2732.6	1826.0	1106.0	
•	128 Se	256 auenc	512 e Lenc	1024 ath		128 Se	256 auenc	512 e Lenc	1024 ath	
	2 - 4 - 8 - 16 - 32 - 64 - 128 - 1 - 2 - 4 - 8 - 16 - 32 - 64 -	1 - 0.8 2 - 3.2 4 - 12.5 8 - 40.8 16 - 152.4 32 - 524.7 64 - 1628.6 128 - 4619.3 See AW 1 - 1.1 2 - 4.3 4 - 16.5 8 - 60.8 16 - 208.4 32 - 587.2 64 - 1503.3 128 - 3721.9	1 - 0.8 0.8 2 - 3.2 3.2 4 - 12.5 12.3 8 - 40.8 39.8 16 - 152.4 143.6 32 - 524.7 474.0 64 - 1628.6 1405.6 128 - 4619.3 3715.6 Sequence AWQ CC 1 - 1.1 1.1 2 - 4.3 4.2 4 - 16.5 15.8 8 - 60.8 56.2 16 - 208.4 182.3 32 - 587.2 487.0 64 - 1503.3 1198.2 128 - 3721.9 2809.7 128 256	1 - 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8	1 - 0.8	1 - 0.8	1 - 0.8	1 - 0.8	1 - 0.8	