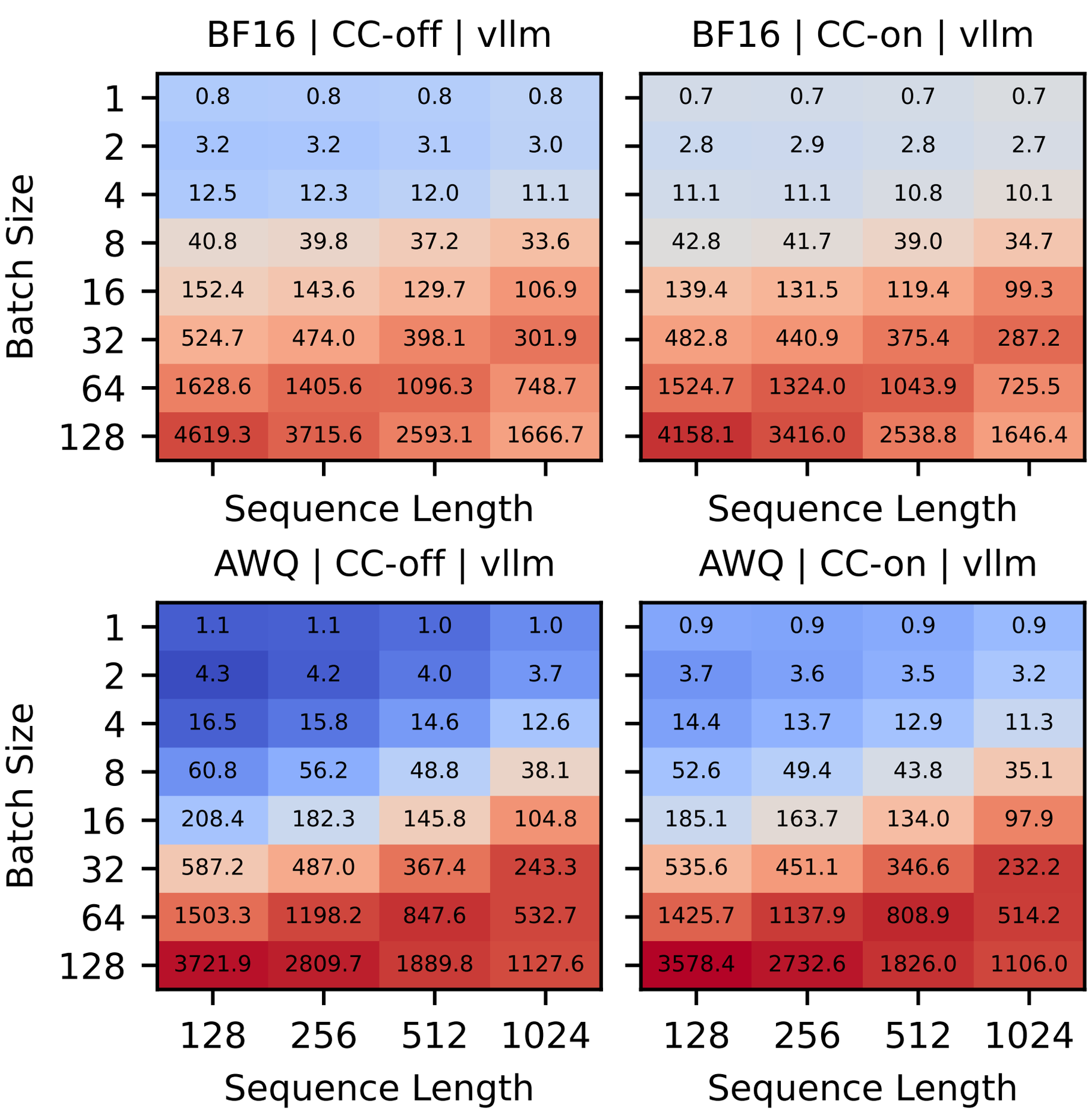
BF16 | CC-off | vllm
BF16 | CC-on | vllm
AWQ | CC-off | vllm
AWQ | CC-on | vllm
Batch Size
Sequence Length
1
2
4
8
16
32
64
128
128
256
512
1024
0.8 0.8 0.8 0.8
3.2 3.2 3.1 3.0
12.5 12.3 12.0 11.1
40.8 39.8 37.2 33.6
152.4 143.6 129.7 106.9
524.7 474.0 398.1 301.9
1628.6 1405.6 1096.3 748.7
4619.3 3715.6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