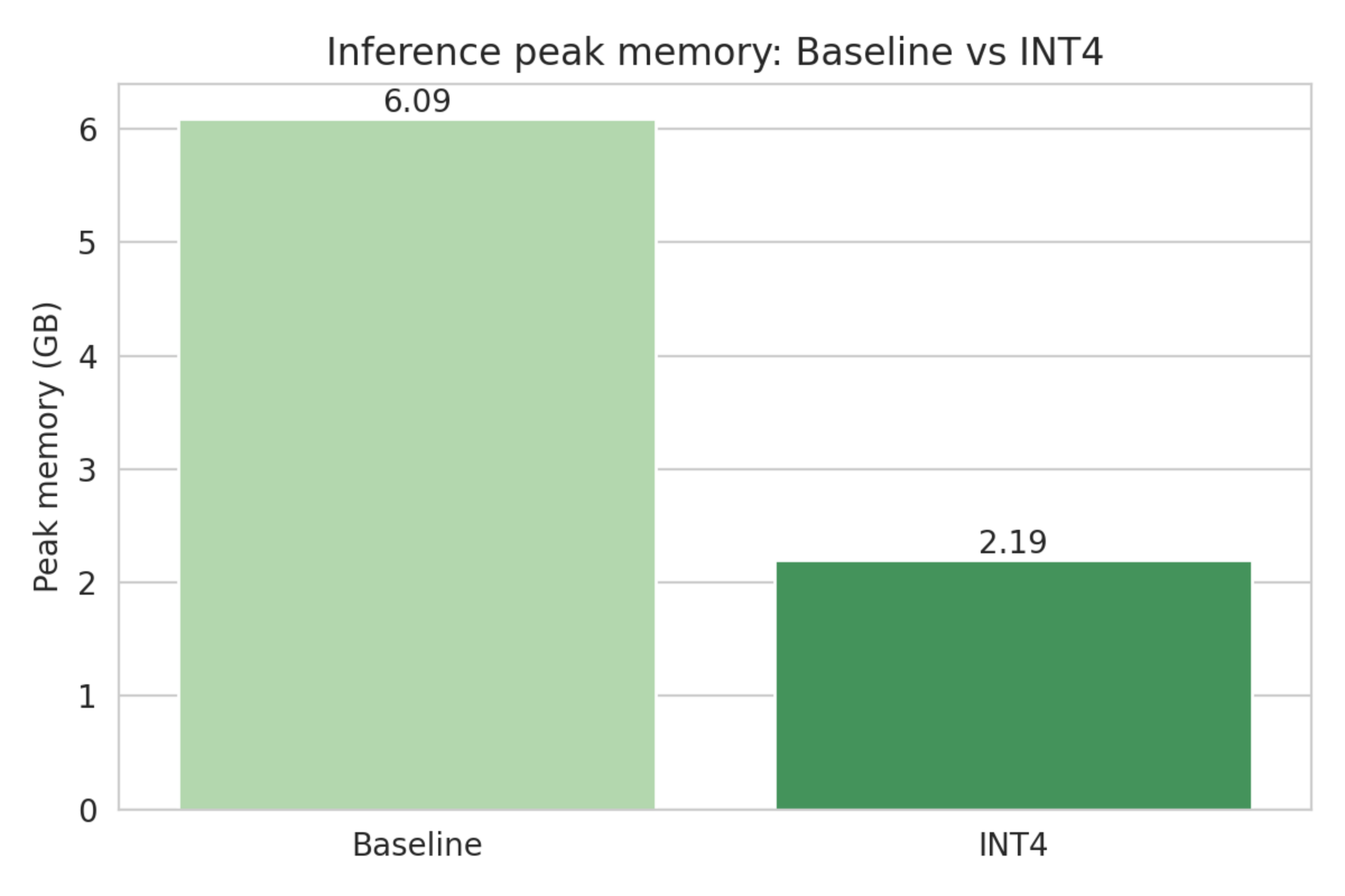
Inference peak memory: Baseline vs INT4
Peak memory (GB)
0
1
2
3
4
5
6
6.09
2.19
Baseline
INT4