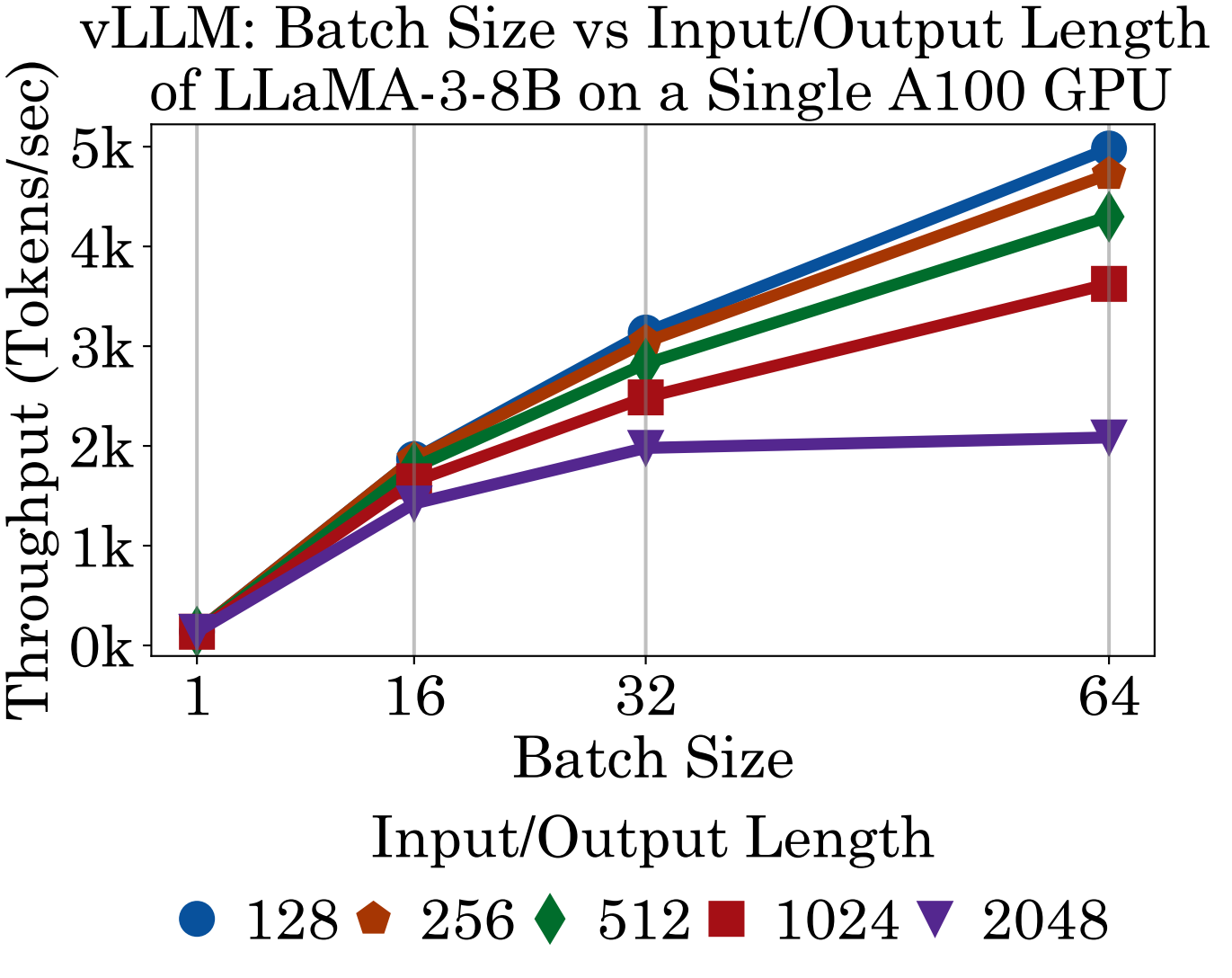
vLLM: Batch Size vs Input/Output Length
of LLaMA-3-8B on a Single A100 GPU
Throughput (Tokens/sec)
0k
1k
2k
3k
4k
5k
1
16
32
64
Batch Size
Input/Output Length
128
256
512
1024
2048