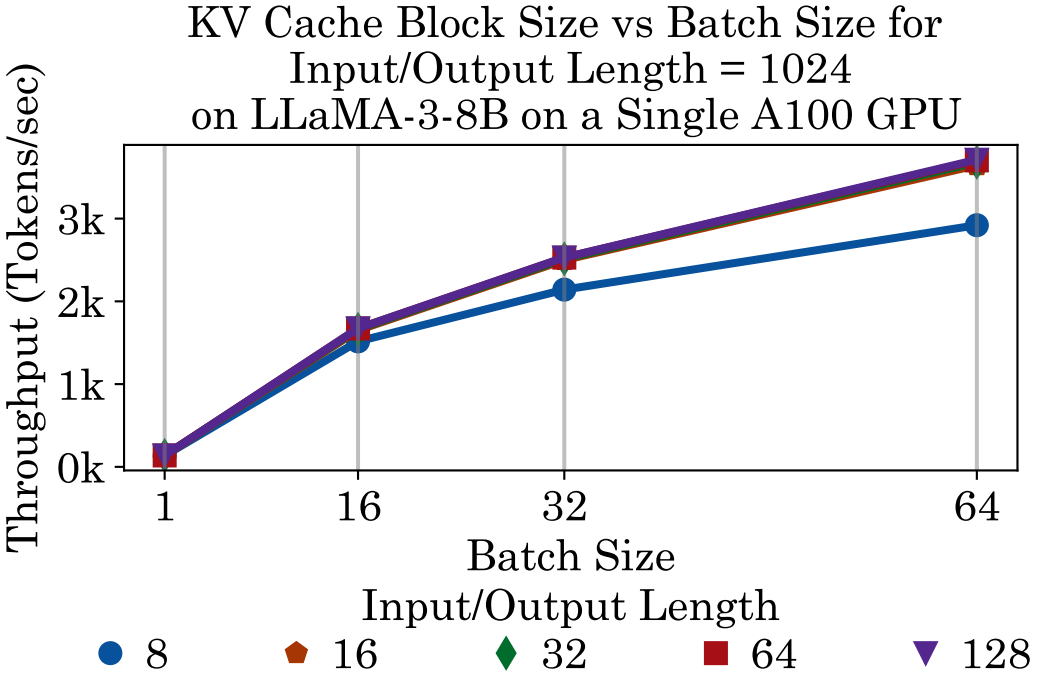
KV Cache Block Size vs Batch Size for
Input/Output Length = 1024
on LLaMA-3-8B on a Single A100 GPU
Throughput (Tokens/sec)
0k
1k
2k
3k
1
16
32
64
Batch Size
Input/Output Length
8
16
32
64
128