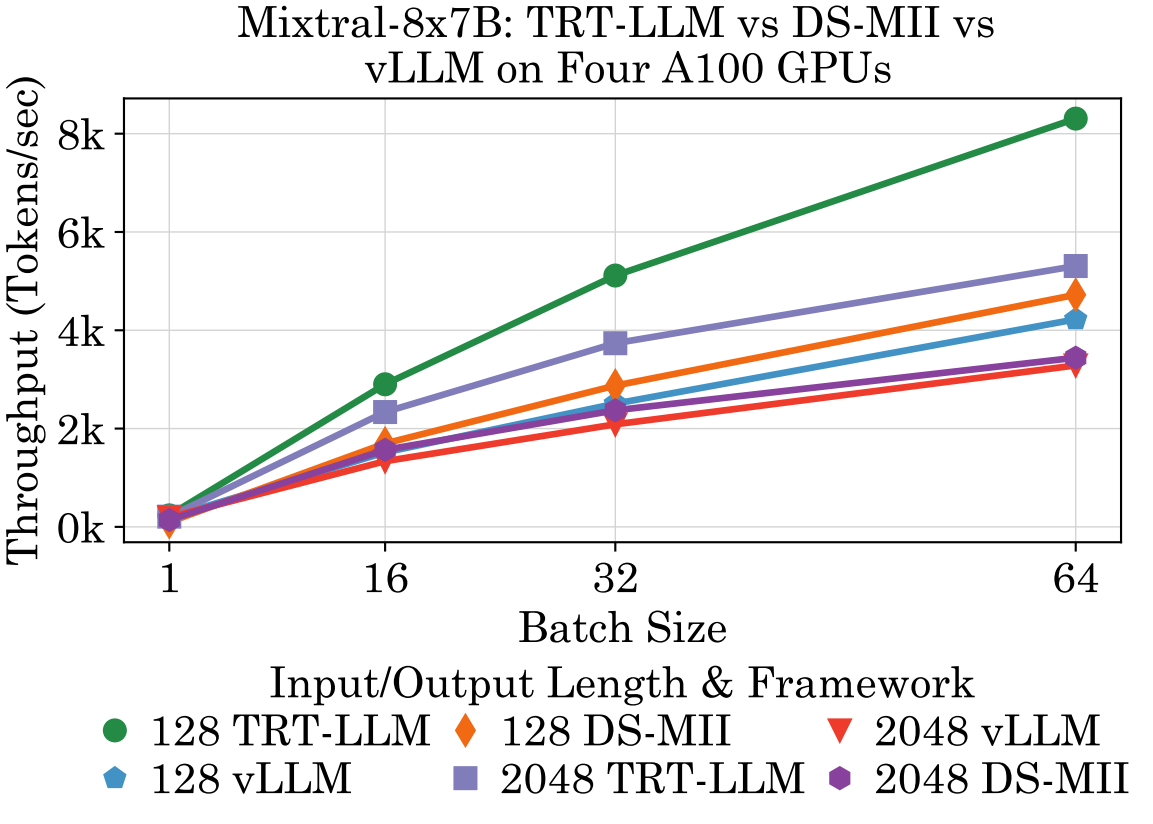
Mixtral-8x7B: TRT-LLM vs DS-MII vs vLLM on Four A100 GPUs
Throughput (Tokens/sec)
0k
2k
4k
6k
8k
1
16
32
64
Batch Size
Input/Output Length & Framework
128 TRT-LLM
128 DS-MII
2048 vLLM
128 vLLM
2048 TRT-LLM
2048 DS-MII