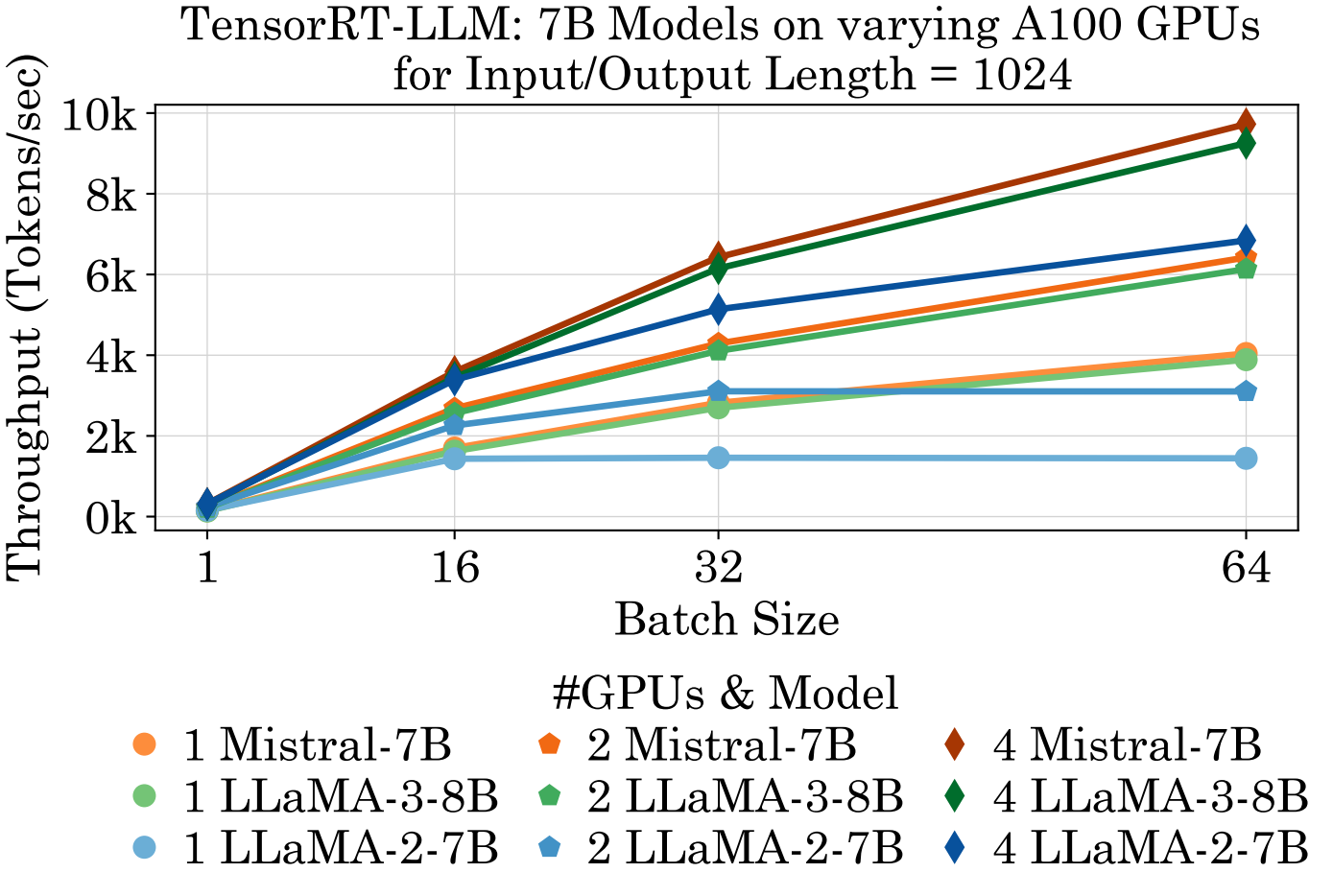
TensorRT-LLM: 7B Models on varying A100 GPUs
for Input/Output Length = 1024
Throughput (Tokens/sec)
10k
8k
6k
4k
2k
0k
1
16
32
64
Batch Size
#GPUs & Model
1 Mistral-7B
2 Mistral-7B
4 Mistral-7B
1 LLaMA-3-8B
2 LLaMA-3-8B
4 LLaMA-3-8B
1 LLaMA-2-7B
2 LLaMA-2-7B
4 LLaMA-2-7B