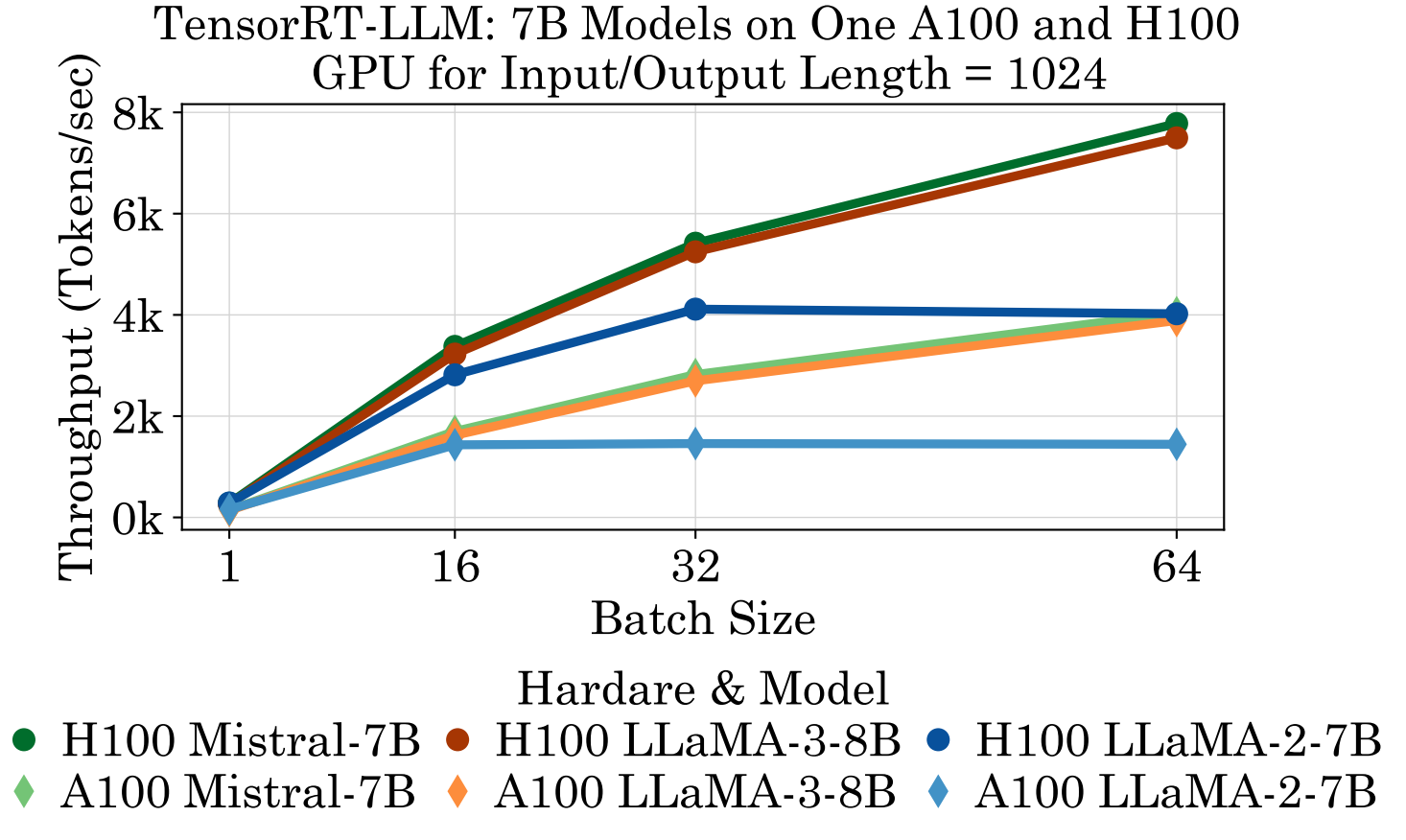
TensorRT-LLM: 7B Models on One A100 and H100 GPU for Input/Output Length = 1024
Throughput (Tokens/sec)
8k
6k
4k
2k
0k
1
16
32
64
Batch Size
Hardare & Model
H100 Mistral-7B
H100 LLaMA-3-8B
H100 LLaMA-2-7B
A100 Mistral-7B
A100 LLaMA-3-8B
A100 LLaMA-2-7B