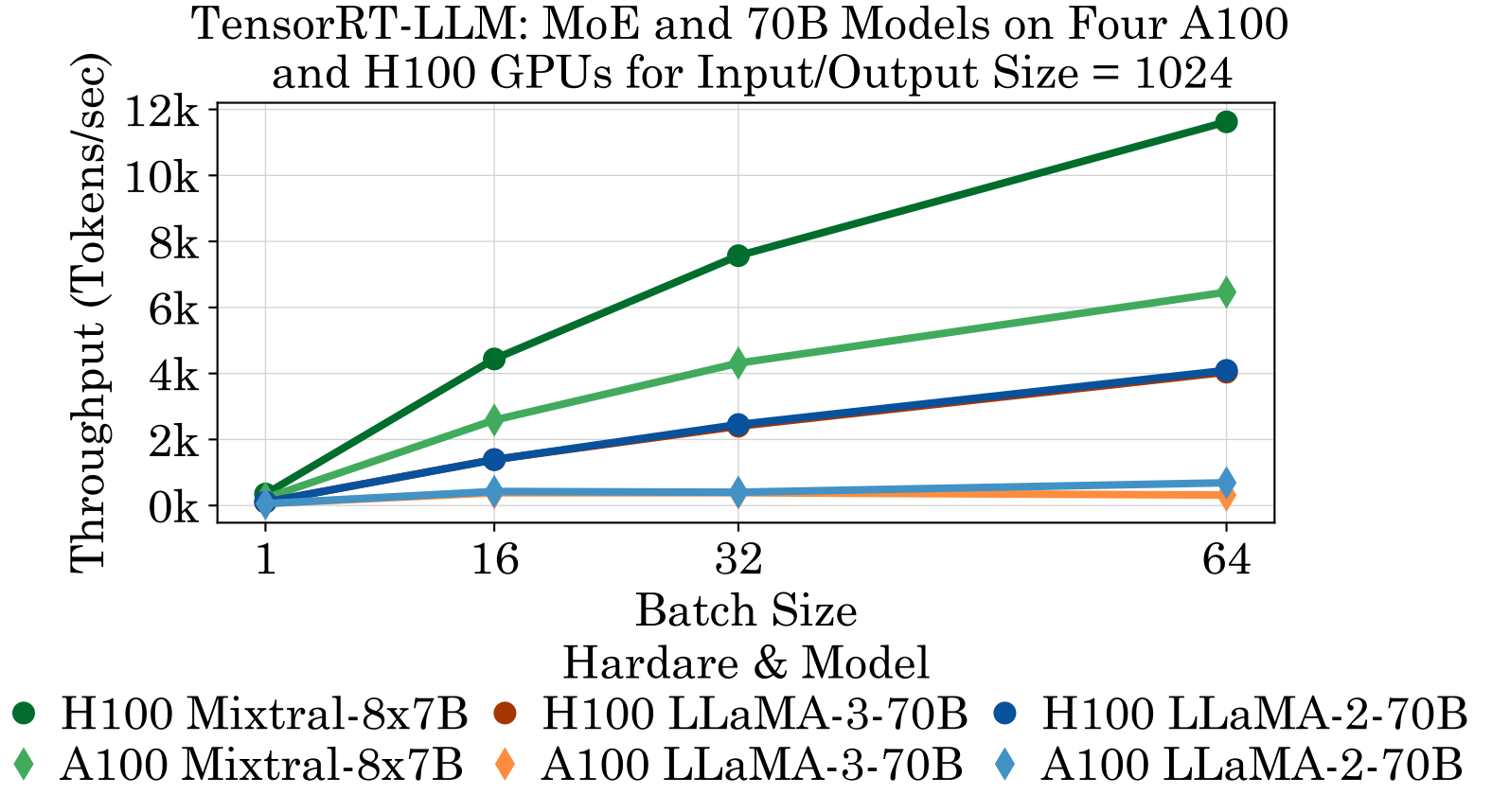
TensorRT-LLM: MoE and 70B Models on Four A100 and H100 GPUs for Input/Output Size = 1024
Throughput (Tokens/sec)
12k
10k
8k
6k
4k
2k
0k
1
16
32
64
Batch Size
Hardare & Model
H100 Mixtral-8x7B
H100 LLaMA-3-70B
H100 LLaMA-2-70B
A100 Mixtral-8x7B
A100 LLaMA-3-70B
A100 LLaMA-2-70B