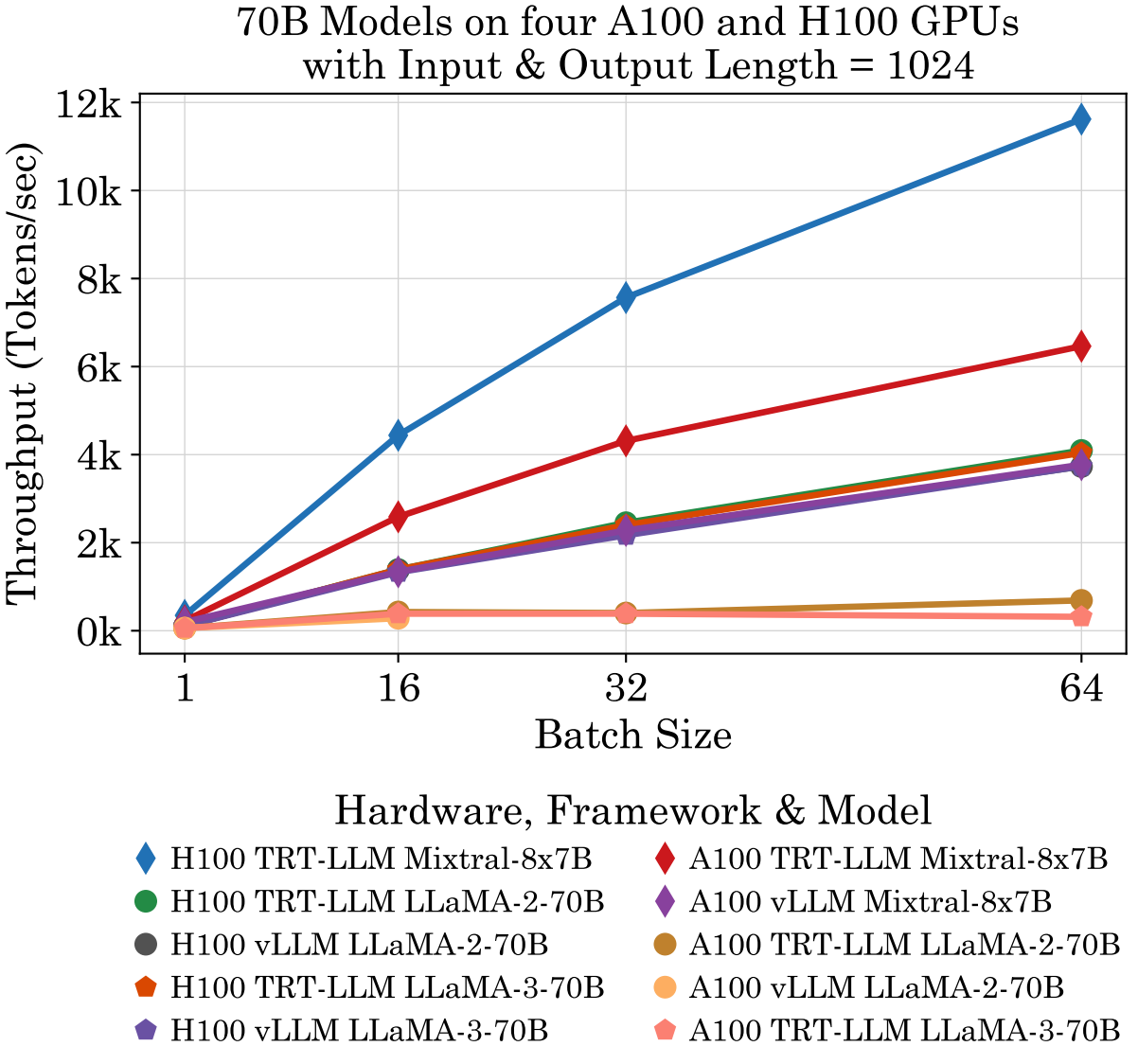
70B Models on four A100 and H100 GPUs
with Input & Output Length = 1024
Throughput (Tokens/sec)
12k
10k
8k
6k
4k
2k
0k
1
16
32
64
Batch Size
Hardware, Framework & Model
H100 TRT-LLM Mixtral-8x7B
H100 TRT-LLM LLaMA-2-70B
H100 vLLM LLaMA-2-70B
H100 TRT-LLM LLaMA-3-70B
H100 vLLM LLaMA-3-70B
A100 TRT-LLM Mixtral-8x7B
A100 vLLM Mixtral-8x7B
A100 TRT-LLM LLaMA-2-70B
A100 vLLM LLaMA-2-70B
A100 TRT-LLM LLaMA-3-70B