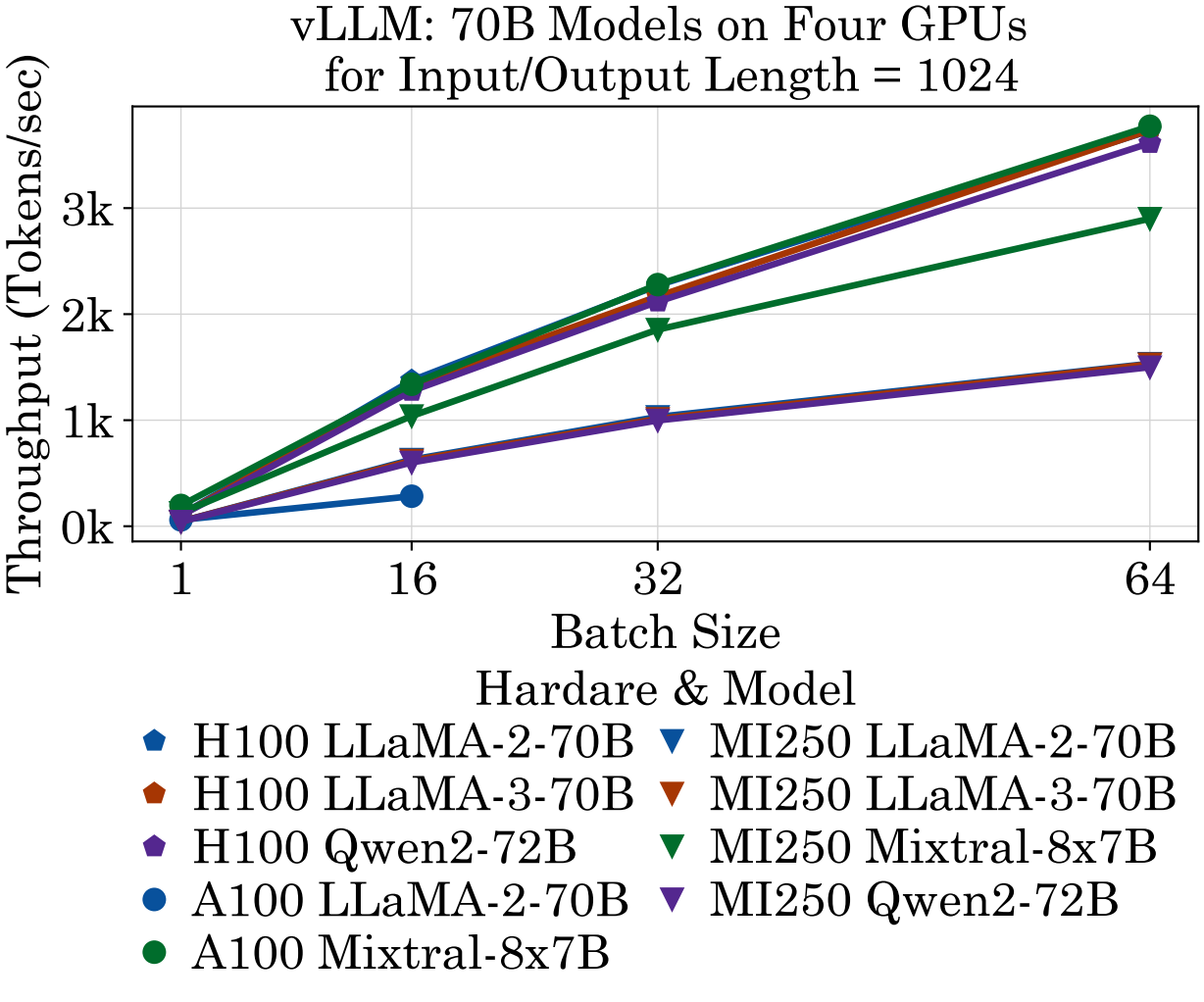
vLLM: 70B Models on Four GPUs
for Input/Output Length = 1024
Throughput (Tokens/sec)
0k
1k
2k
3k
1
16
32
64
Batch Size
Hardare & Model
H100 LLaMA-2-70B
H100 LLaMA-3-70B
H100 Qwen2-72B
A100 LLaMA-2-70B
A100 Mixtral-8x7B
MI250 LLaMA-2-70B
MI250 LLaMA-3-70B
MI250 Mixtral-8x7B
MI250 Qwen2-72B