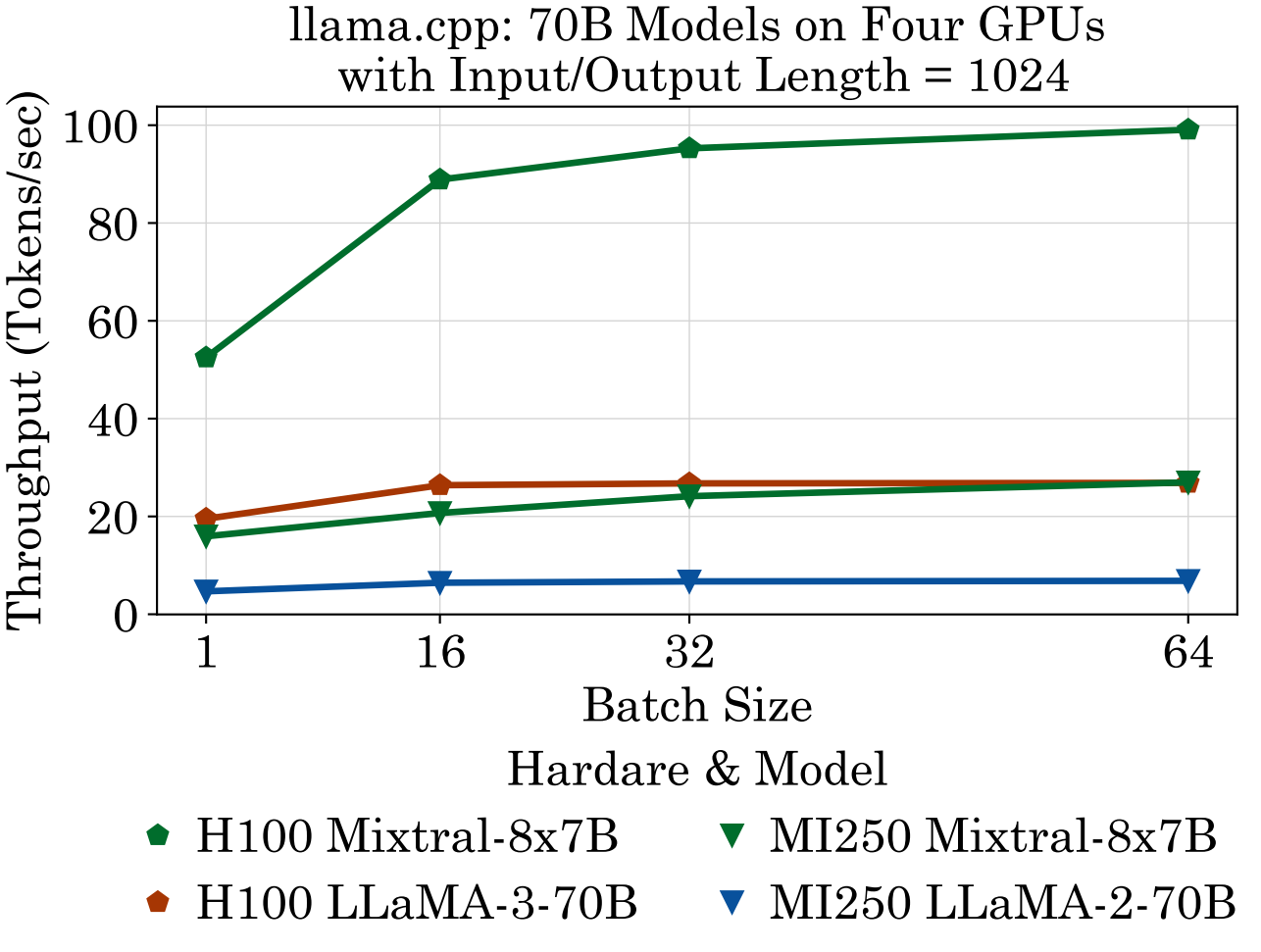
llama.cpp: 70B Models on Four GPUs
with Input/Output Length = 1024
Throughput (Tokens/sec)
0
20
40
60
80
100
1
16
32
64
Batch Size
Hardare & Model
H100 Mixtral-8x7B
H100 LLaMA-3-70B
MI250 Mixtral-8x7B
MI250 LLaMA-2-70B