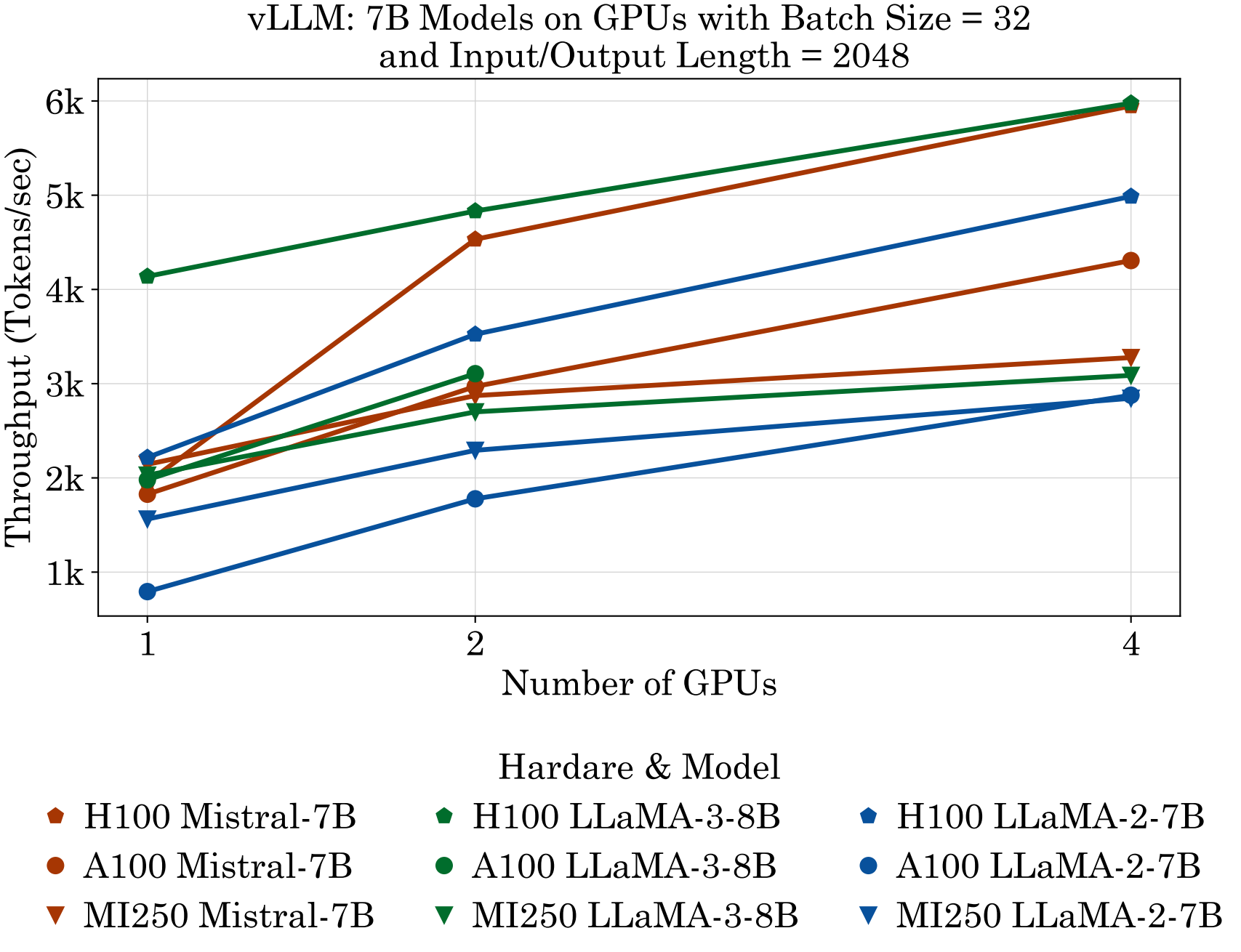
vLLM: 7B Models on GPUs with Batch Size = 32
and Input/Output Length = 2048
Throughput (Tokens/sec)
6k
5k
4k
3k
2k
1k
1
2
4
Number of GPUs
Hardare & Model
H100 Mistral-7B
H100 LLaMA-3-8B
H100 LLaMA-2-7B
A100 Mistral-7B
A100 LLaMA-3-8B
A100 LLaMA-2-7B
MI250 Mistral-7B
MI250 LLaMA-3-8B
MI250 LLaMA-2-7B