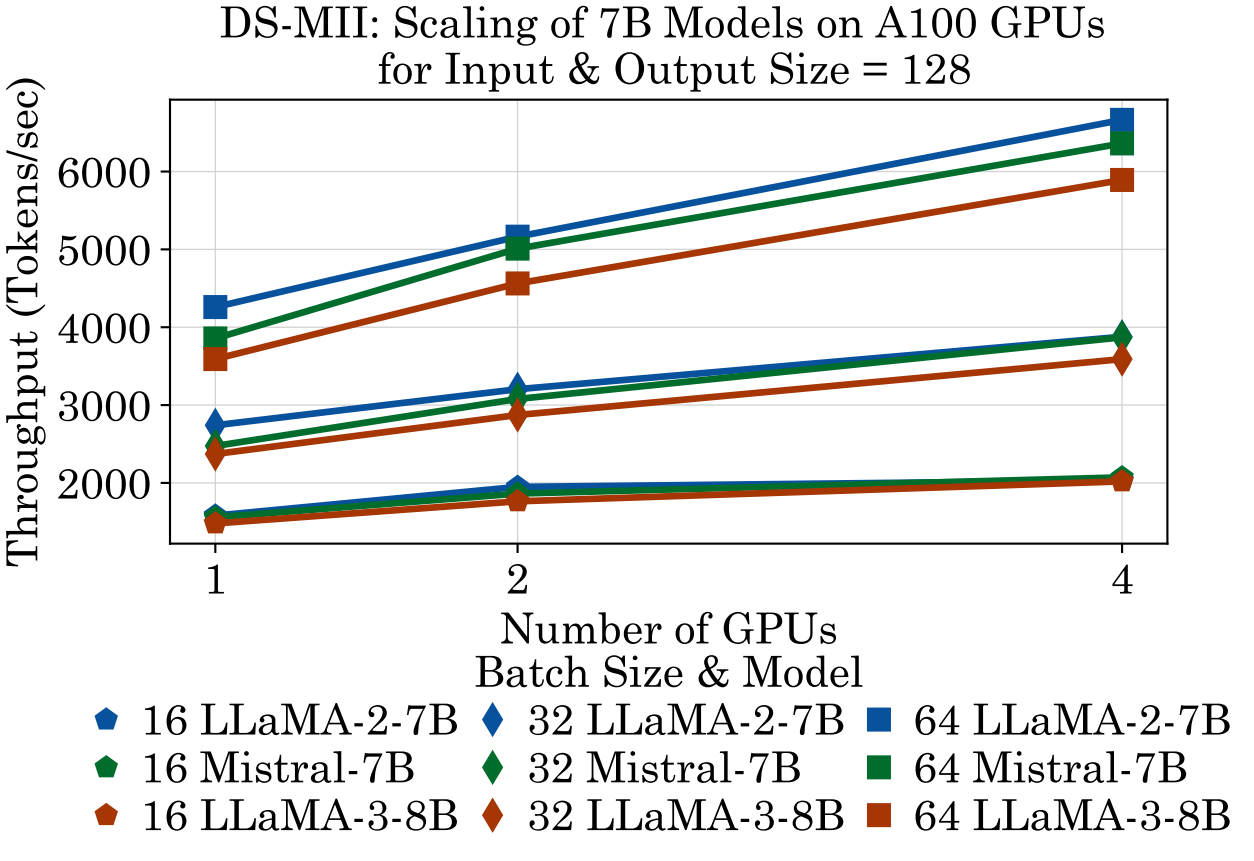
DS-MII: Scaling of 7B Models on A100 GPUs
for Input & Output Size = 128
Throughput (Tokens/sec)
6000
5000
4000
3000
2000
1
2
4
Number of GPUs
Batch Size & Model
16 LLaMA-2-7B
32 LLaMA-2-7B
64 LLaMA-2-7B
16 Mistral-7B
32 Mistral-7B
64 Mistral-7B
16 LLaMA-3-8B
32 LLaMA-3-8B
64 LLaMA-3-8B