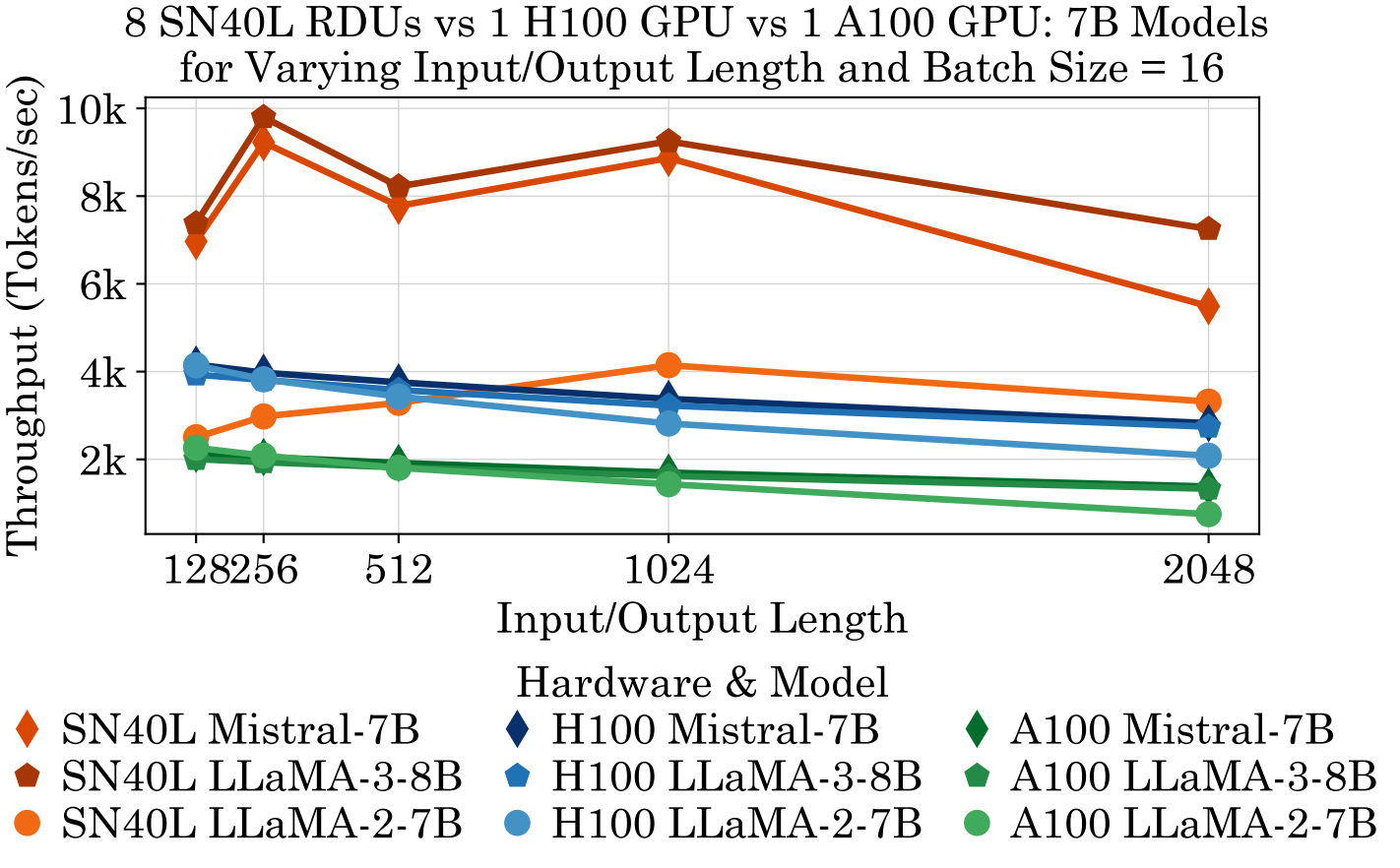
8 SN40L RDUs vs 1 H100 GPU vs 1 A100 GPU: 7B Models
for Varying Input/Output Length and Batch Size = 16
Throughput (Tokens/sec)
10k
8k
6k
4k
2k
128
256
512
1024
2048
Input/Output Length
Hardware & Model
SN40L Mistral-7B
SN40L LLaMA-3-8B
SN40L LLaMA-2-7B
H100 Mistral-7B
H100 LLaMA-3-8B
H100 LLaMA-2-7B
A100 Mistral-7B
A100 LLaMA-3-8B
A100 LLaMA-2-7B