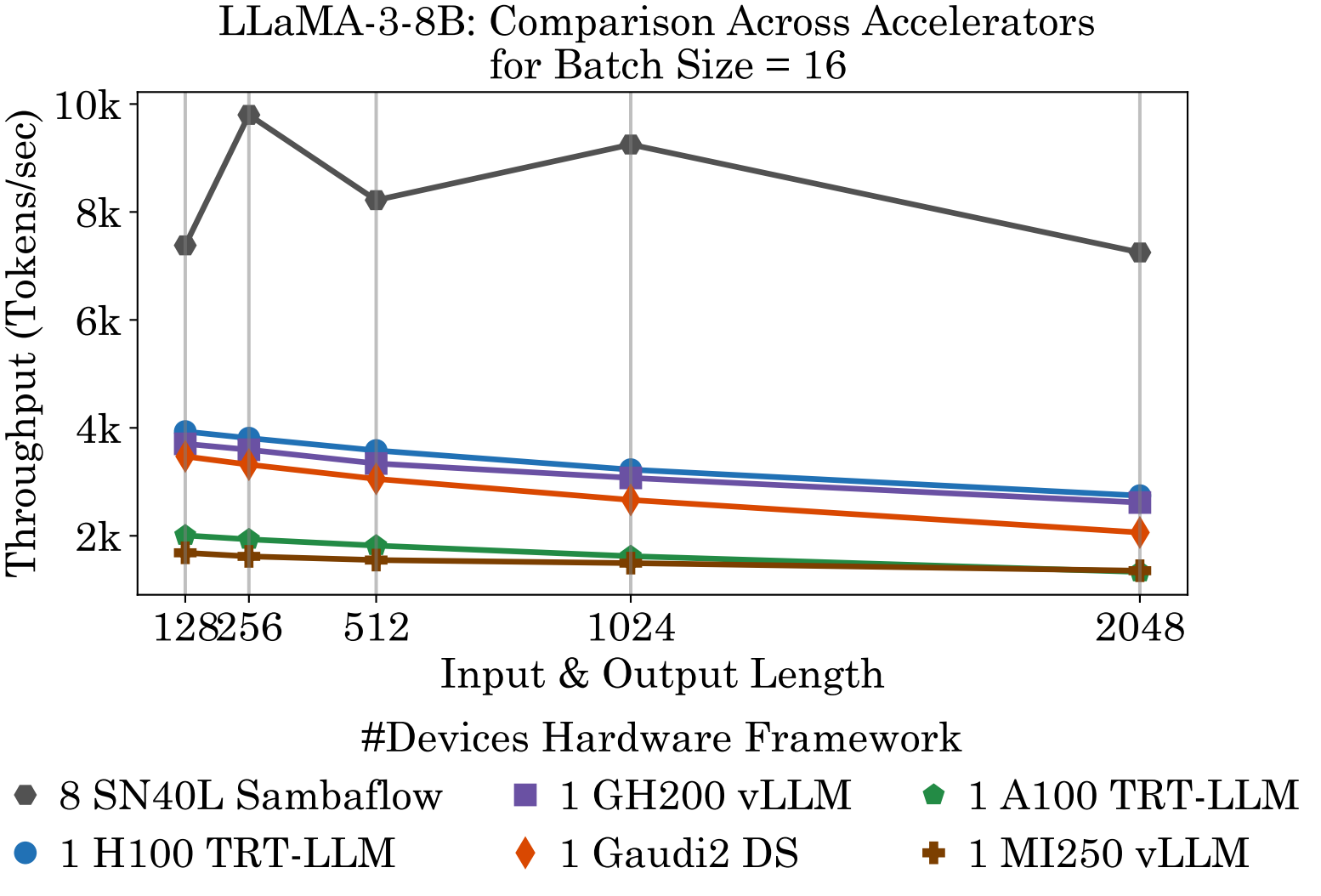
LLaMA-3-8B: Comparison Across Accelerators
for Batch Size = 16
Throughput (Tokens/sec)
10k
8k
6k
4k
2k
128
256
512
1024
2048
Input & Output Length
#Devices Hardware Framework
8 SN40L Sambaflow
1 GH200 vLLM
1 A100 TRT-LLM
1 H100 TRT-LLM
1 Gaudi2 DS
1 MI250 vLLM