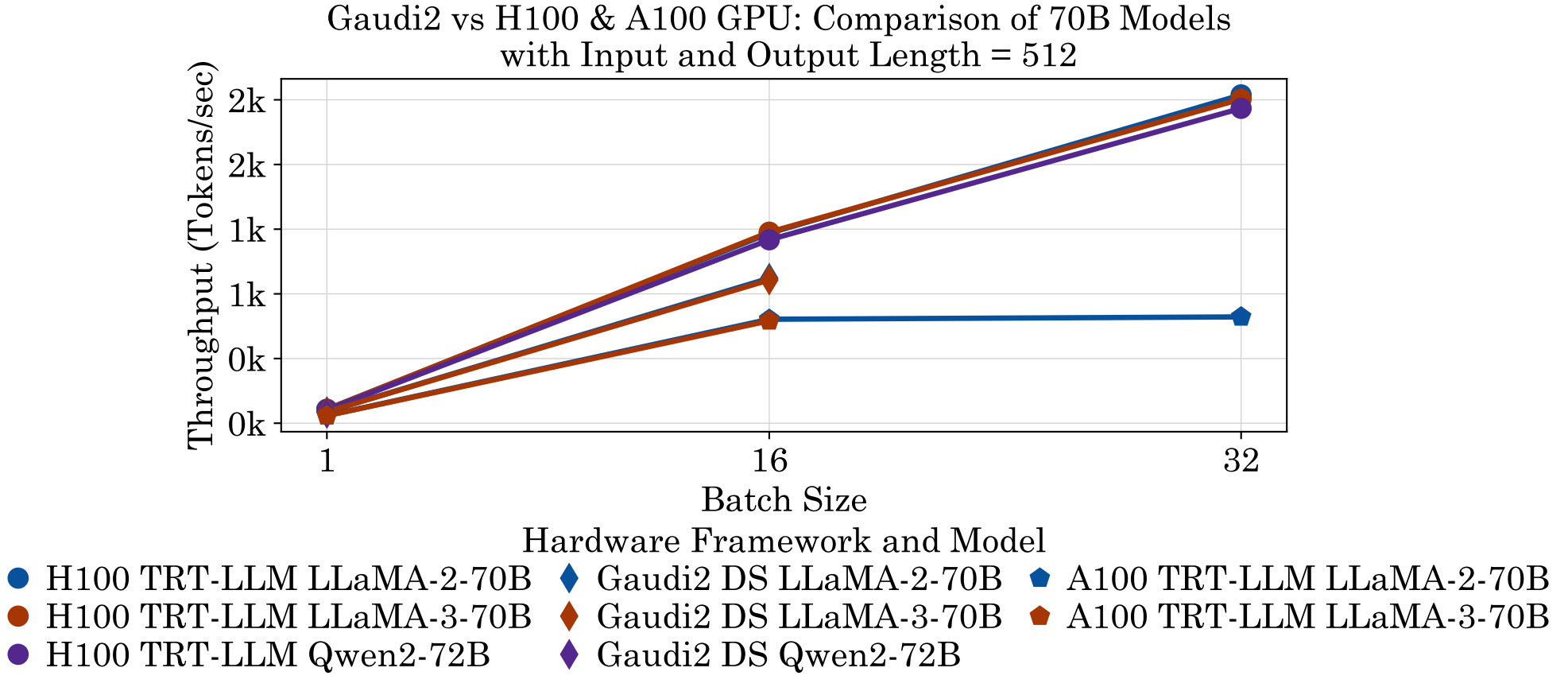
Gaudi2 vs H100 & A100 GPU: Comparison of 70B Models
with Input and Output Length = 512
Throughput (Tokens/sec)
2k
2k
1k
1k
0k
0k
1
16
32
Batch Size
Hardware Framework and Model
H100 TRT-LLM LLaMA-2-70B
H100 TRT-LLM LLaMA-3-70B
H100 TRT-LLM Qwen2-72B
Gaudi2 DS LLaMA-2-70B
Gaudi2 DS LLaMA-3-70B
Gaudi2 DS Qwen2-72B
A100 TRT-LLM LLaMA-2-70B
A100 TRT-LLM LLaMA-3-70B