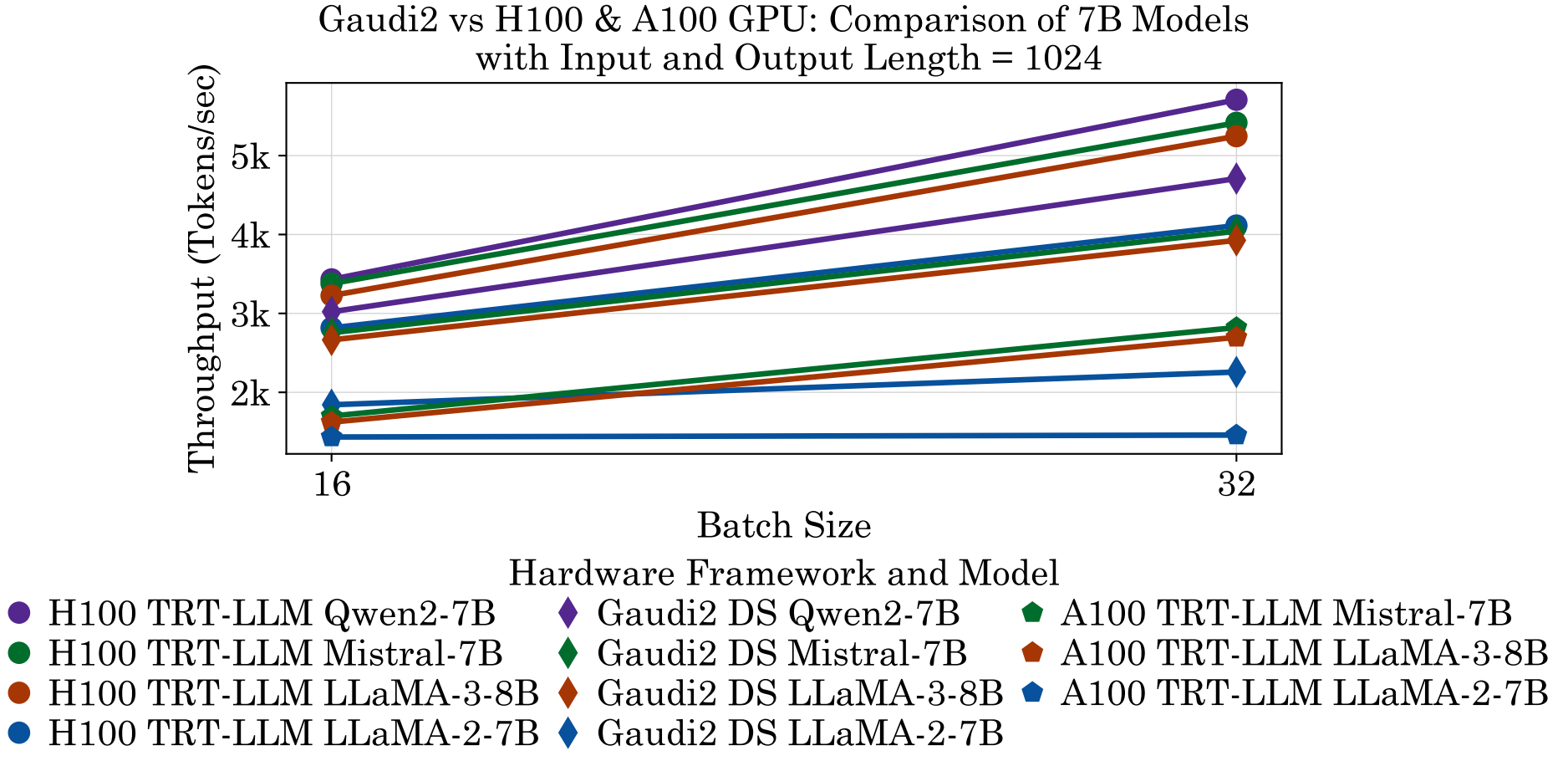
Gaudi2 vs H100 & A100 GPU: Comparison of 7B Models
with Input and Output Length = 1024
Throughput (Tokens/sec)
5k
4k
3k
2k
16
32
Batch Size
Hardware Framework and Model
H100 TRT-LLM Qwen2-7B
H100 TRT-LLM Mistral-7B
H100 TRT-LLM LLaMA-3-8B
H100 TRT-LLM LLaMA-2-7B
Gaudi2 DS Qwen2-7B
Gaudi2 DS Mistral-7B
Gaudi2 DS LLaMA-3-8B
Gaudi2 DS LLaMA-2-7B
A100 TRT-LLM Mistral-7B
A100 TRT-LLM LLaMA-3-8B
A100 TRT-LLM LLaMA-2-7B