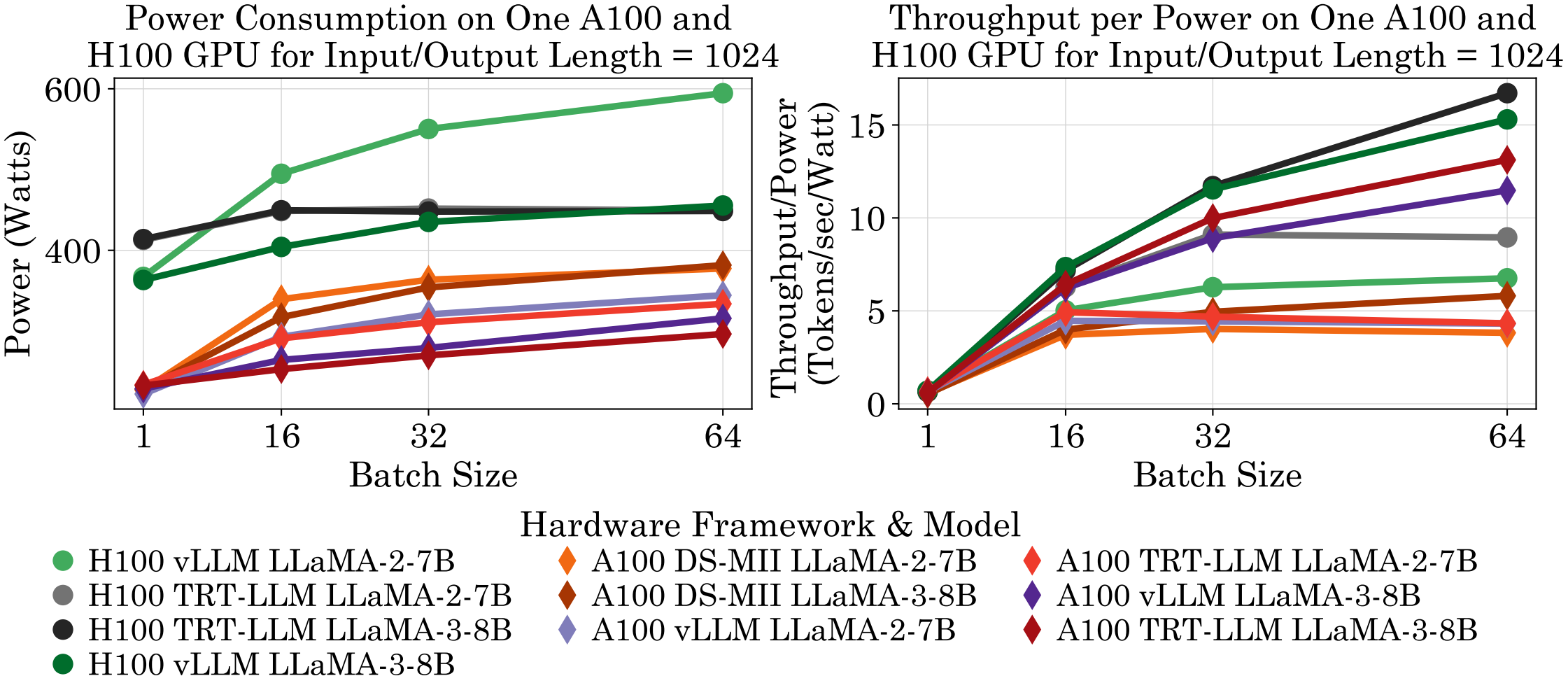
Power Consumption on One A100 and H100 GPU for Input/Output Length = 1024
Power (Watts)
600
400
1
16
32
64
Batch Size
Throughput per Power on One A100 and H100 GPU for Input/Output Length = 1024
Throughput/Power (Tokens/sec/Watt)
15
10
5
0
1
16
32
64
Batch Size
Hardware Framework & Model
H100 vLLM LLaMA-2-7B
H100 TRT-LLM LLaMA-2-7B
H100 TRT-LLM LLaMA-3-8B
H100 vLLM LLaMA-3-8B
A100 DS-MII LLaMA-2-7B
A100 DS-MII LLaMA-3-8B
A100 vLLM LLaMA-2-7B
A100 TRT-LLM LLaMA-2-7B
A100 vLLM LLaMA-3-8B
A100 TRT-LLM LLaMA-3-8B