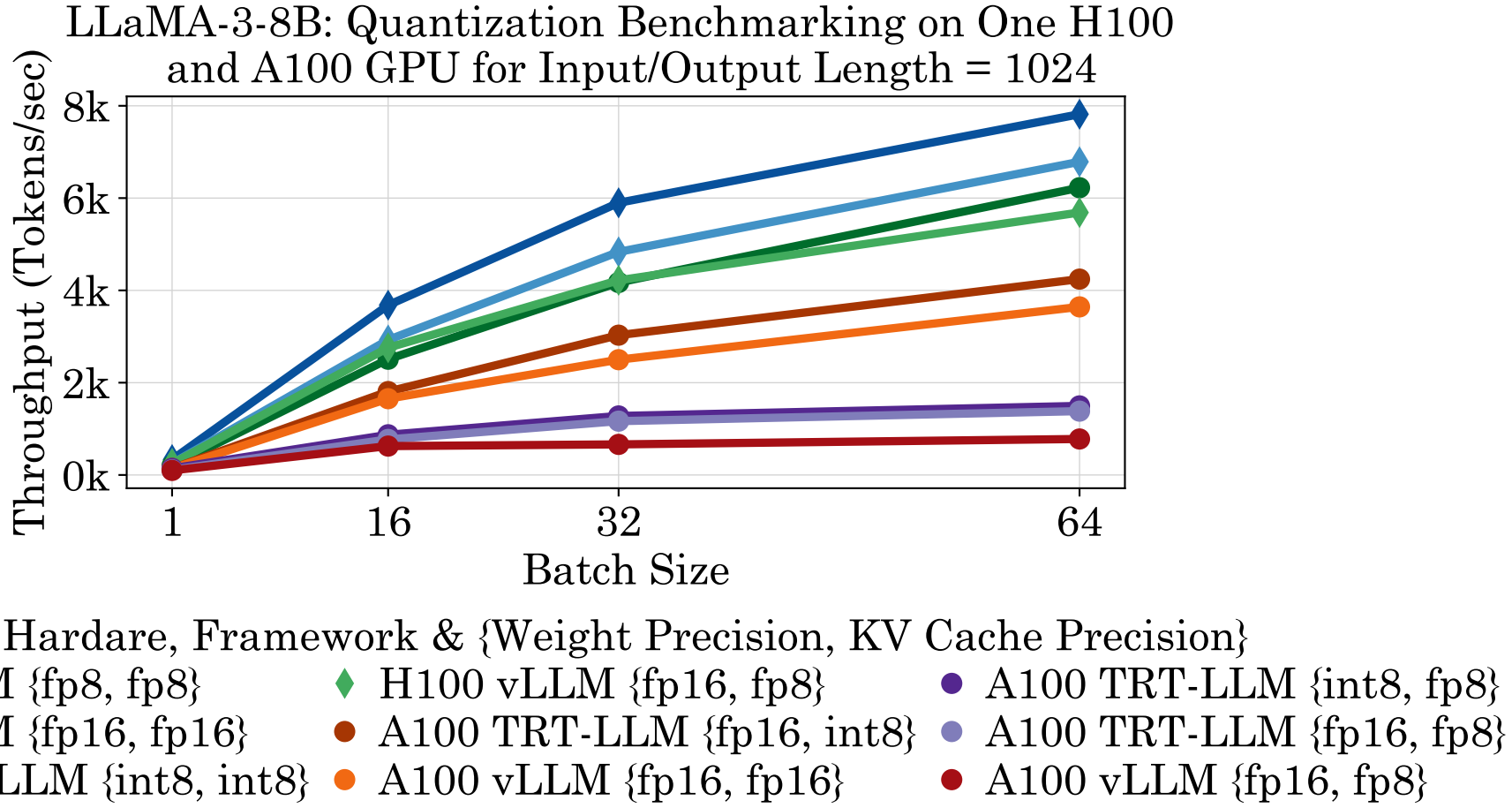
LLaMA-3-8B: Quantization Benchmarking on One H100
and A100 GPU for Input/Output Length = 1024
Throughput (Tokens/sec)
8k
6k
4k
2k
0k
1
16
32
64
Batch Size
Hardare, Framework & {Weight Precision, KV Cache Precision}
H100 vLLM {fp8, fp8}
H100 vLLM {fp16, fp16}
A100 TRT-LLM {int8, int8}
H100 vLLM {fp16, fp8}
A100 TRT-LLM {fp16, int8}
A100 vLLM {fp16, fp16}
A100 TRT-LLM {int8, fp8}
A100 TRT-LLM {fp16, fp8}
A100 vLLM {fp16, fp8}