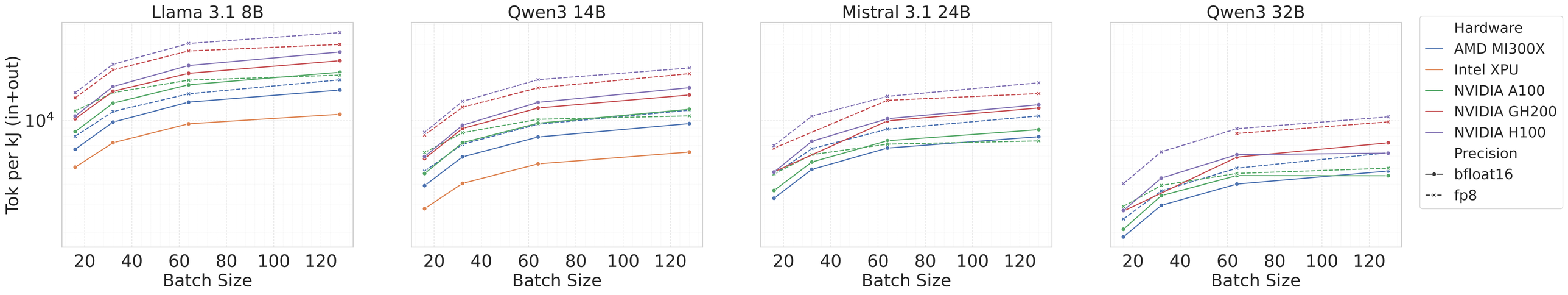
Llama 3.1 8B
Qwen3 14B
Mistral 3.1 24B
Qwen3 32B
Tok per kJ (in+out)
$10^4$
20
40
60
80
100
120
Batch Size
Hardware
AMD MI300X
Intel XPU
NVIDIA A100
NVIDIA GH200
NVIDIA H100
Precision
bfloat16
fp8