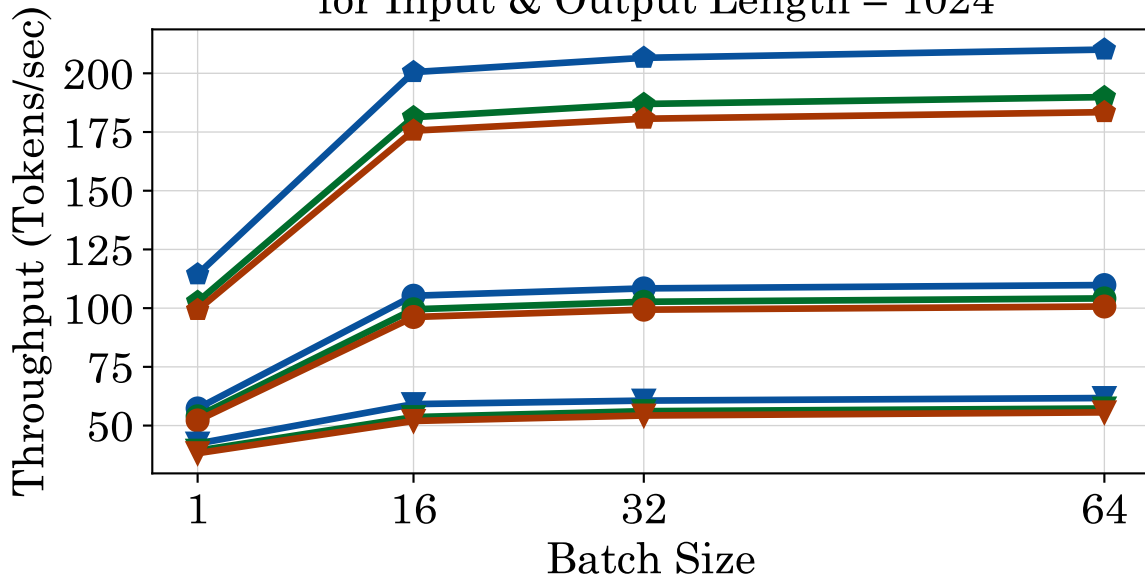


llama.cpp: Performance of 7B Models on One GPU
for Input & Output Length = 1024



Hardare & Model

- | | | |
|-------------------|-------------------|--------------------|
| ◆ H100 LLaMA-2-7B | ● A100 LLaMA-2-7B | ▼ MI250 LLaMA-2-7B |
| ◆ H100 Mistral-7B | ● A100 Mistral-7B | ▼ MI250 Mistral-7B |
| ◆ H100 LLaMA-3-8B | ● A100 LLaMA-3-8B | ▼ MI250 LLaMA-3-8B |