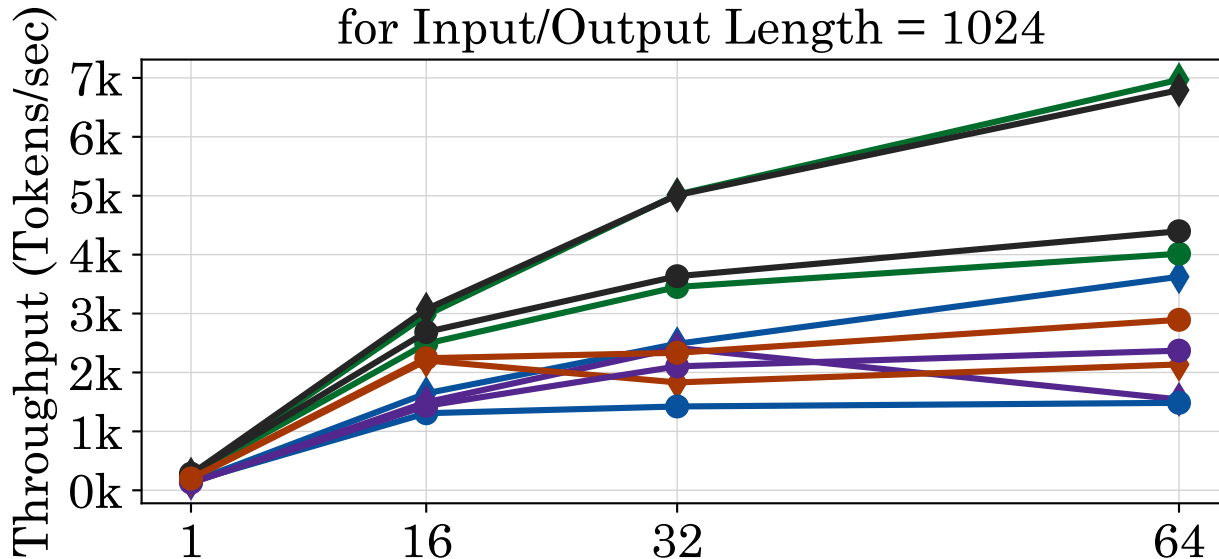


vLLM: 7B Models on One GPU for Input/Output Length = 1024



Batch Size
Hardare & Model

- | | | |
|--------------------|---------------------|---------------------|
| ◆ H100 LLaMA-3-8B | ◆ MI300X LLaMA-3-8B | ● GH200 LLaMA-2-7B |
| ◆ A100 LLaMA-3-8B | ● H100 LLaMA-2-7B | ● MI250 LLaMA-2-7B |
| ◆ GH200 LLaMA-3-8B | ● A100 LLaMA-2-7B | ● MI300X LLaMA-2-7B |
| ◆ MI250 LLaMA-3-8B | | |