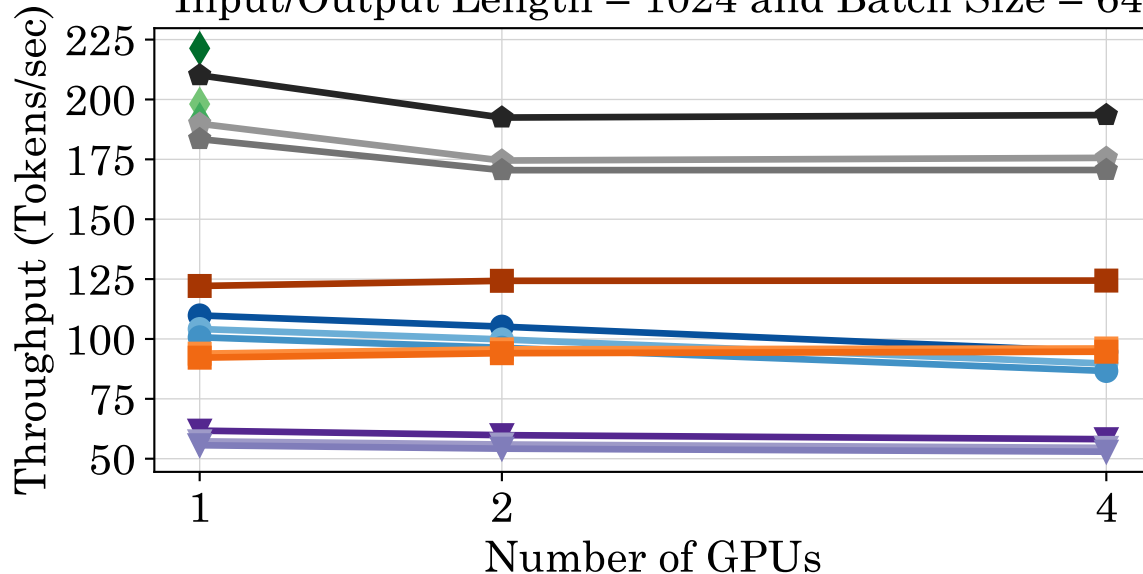


llama.cpp Performance of 7B Models on GPUs for Input/Output Length = 1024 and Batch Size = 64



Hardware & Model

- GH200 LLaMA-2-7B
  H100 LLaMA-3-8B
  MI300X Mistral-7B
- GH200 Mistral-7B
  A100 LLaMA-2-7B
  MI300X LLaMA-3-8B
- GH200 LLaMA-3-8B
  A100 Mistral-7B
  MI250 LLaMA-2-7B
- H100 LLaMA-2-7B
  A100 LLaMA-3-8B
  MI250 Mistral-7B
- H100 Mistral-7B
  MI300X LLaMA-2-7B
  MI250 LLaMA-3-8B