High-Performance
Computing Center
Stuttgart

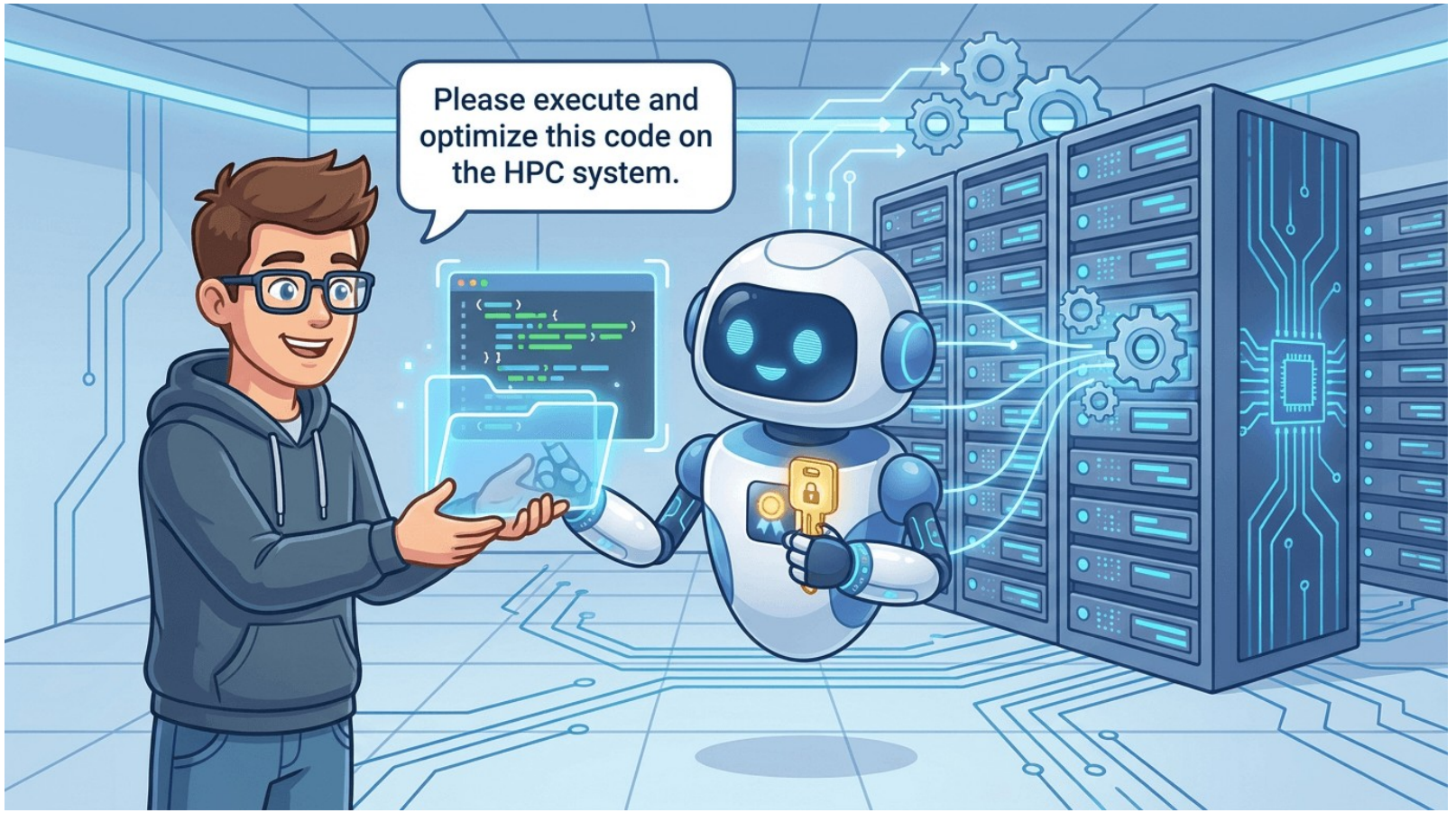
Impact of GPU Virtualization on LLM Inference Performance: A Comparative Study of Bare Metal, vGPU

Qifeng Pan

Outline

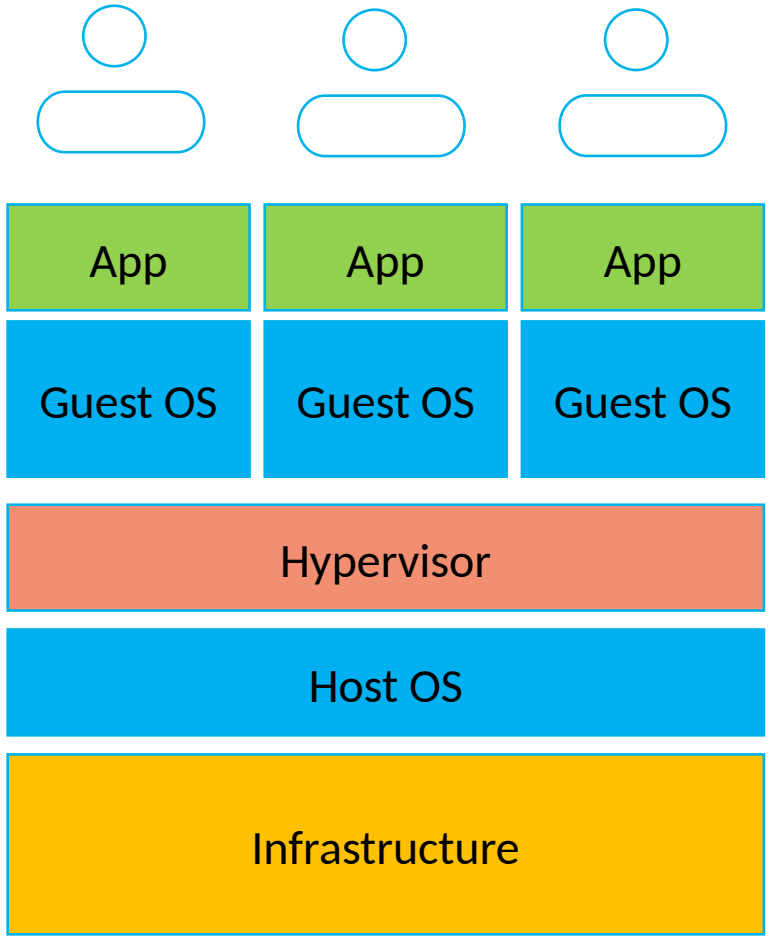
- Background and Motivation
- LLM inferencing and Benchmarking schemes
- Preliminary results
- Future works

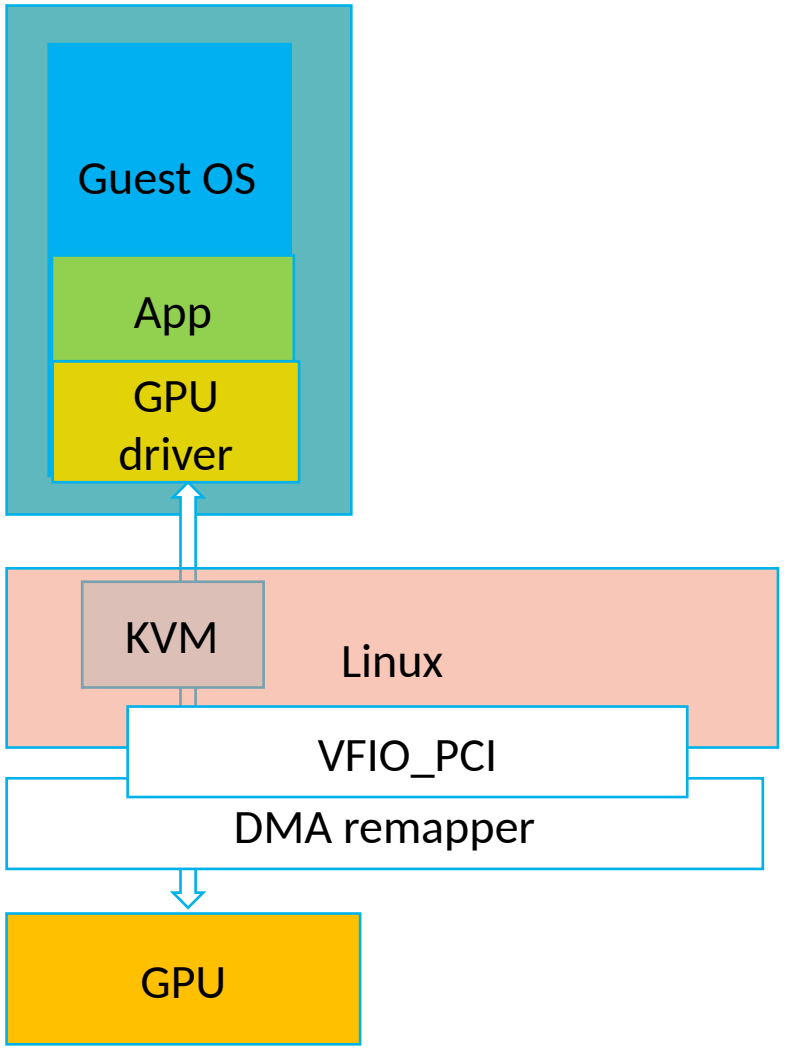
Background & Motivation



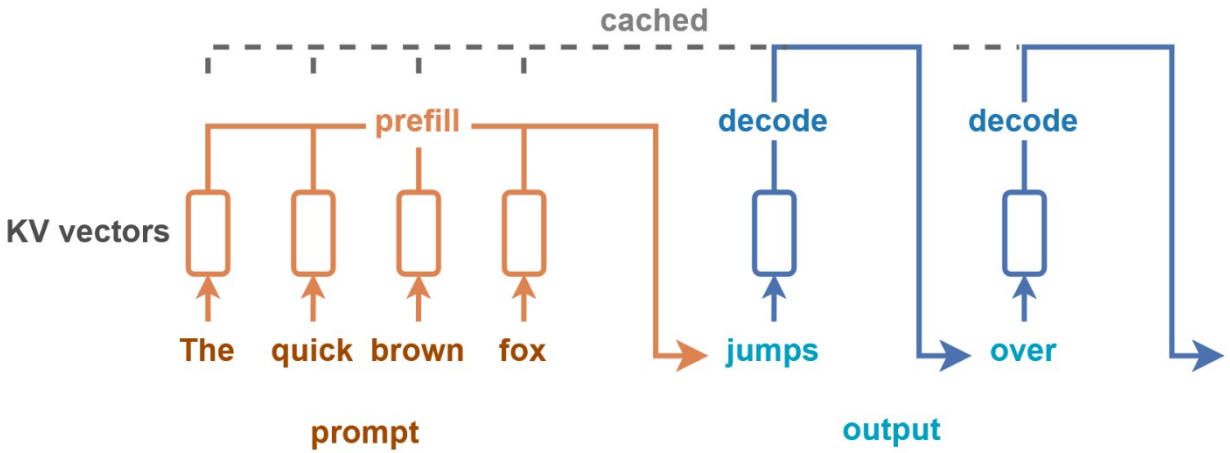
Background & Motivation

Background & Motivation





LLM inferencing



$$Q_{pre} = X_{pre} * W_q \qquad Q_{dec} = X_{dec} * W_q$$

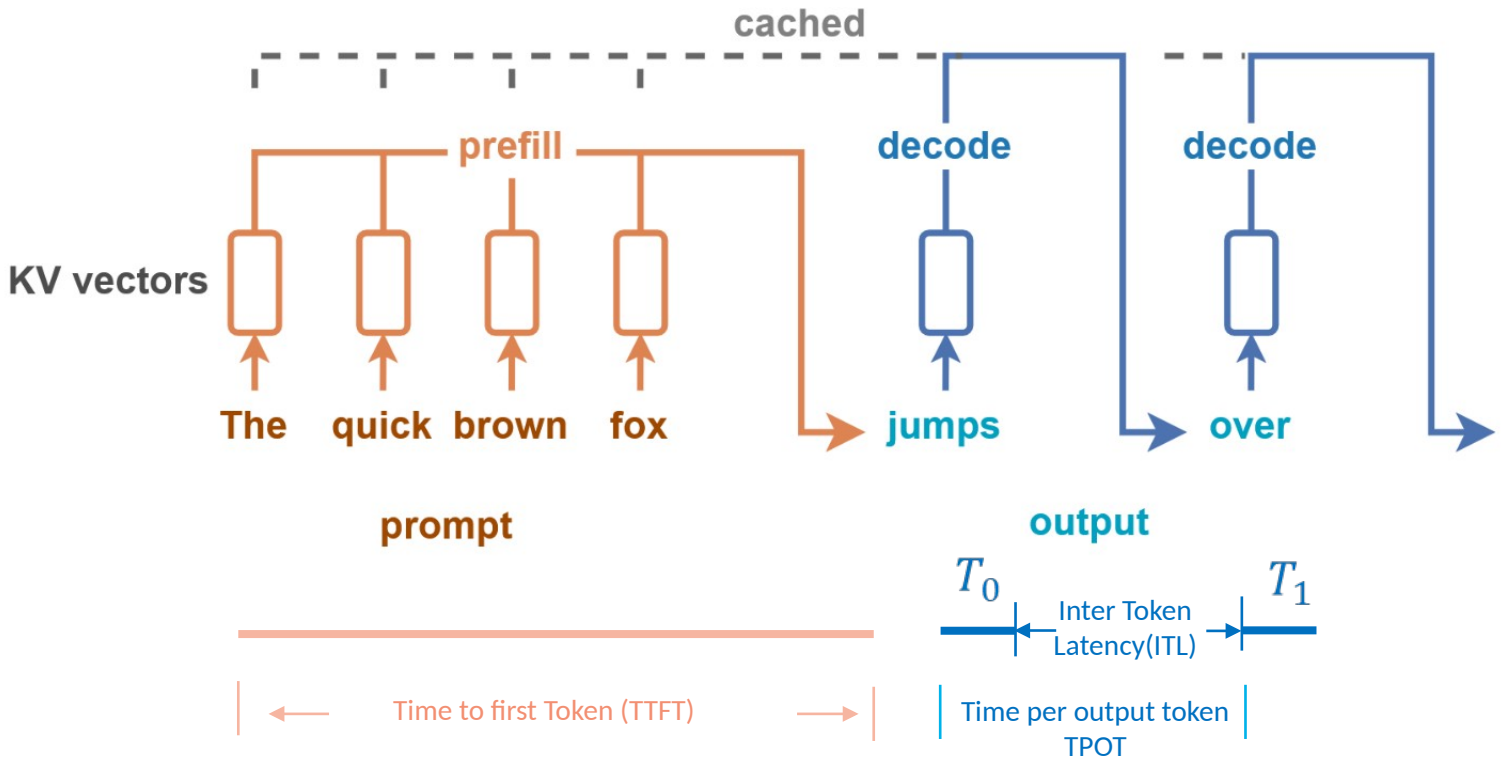
$$K_{pre} = X_{pre} * W_k \quad (n * d) \qquad K_{cat} = [K_{cache}, X_{dec} * W_k] \quad (1 * d)$$

$$V_{pre} = X_{pre} * W_v \qquad V_{cat} = [V_{cache}, X_{dec} * W_v]$$

$$softmax\left(\frac{QK^T}{\sqrt{d}}\right)VW + X$$

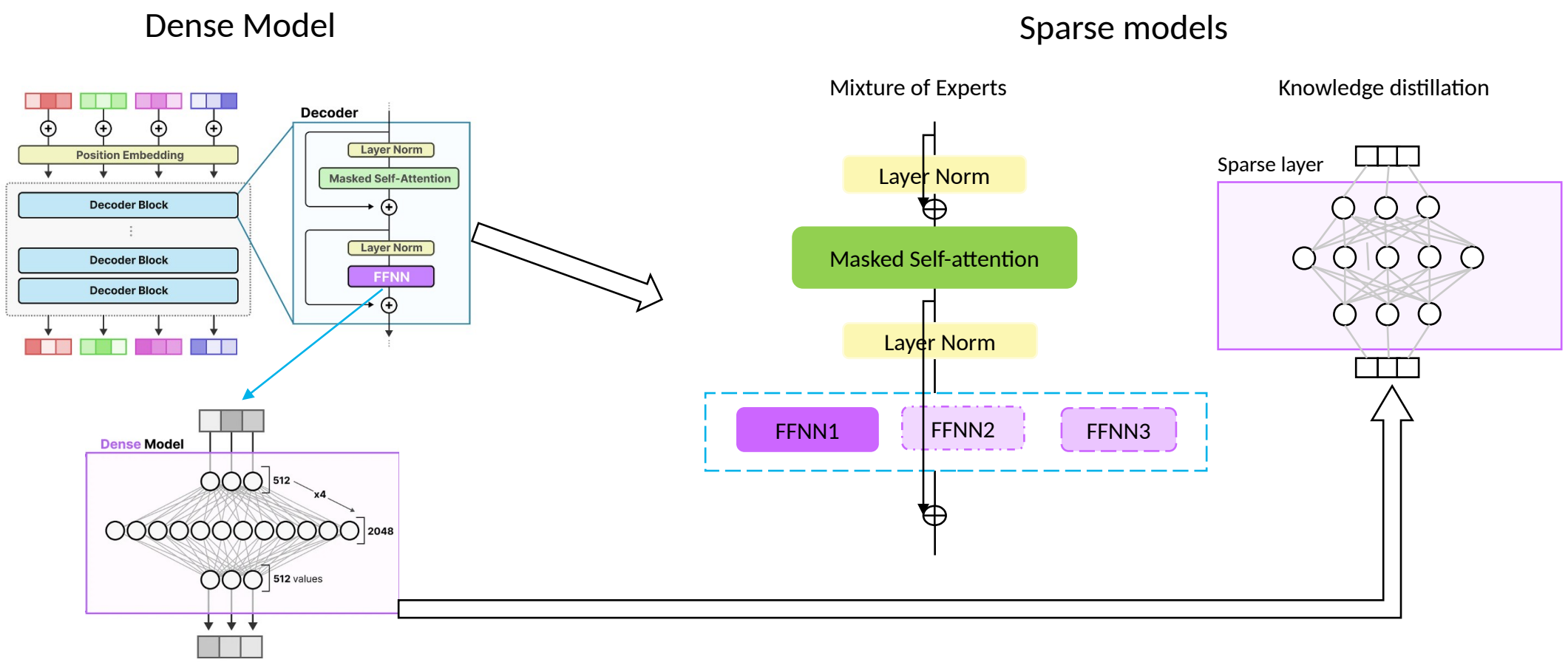
https://huggingface.co/blog/tngtech/llm-performance-prefill-decode-concurrent-requests

LLM inferencing



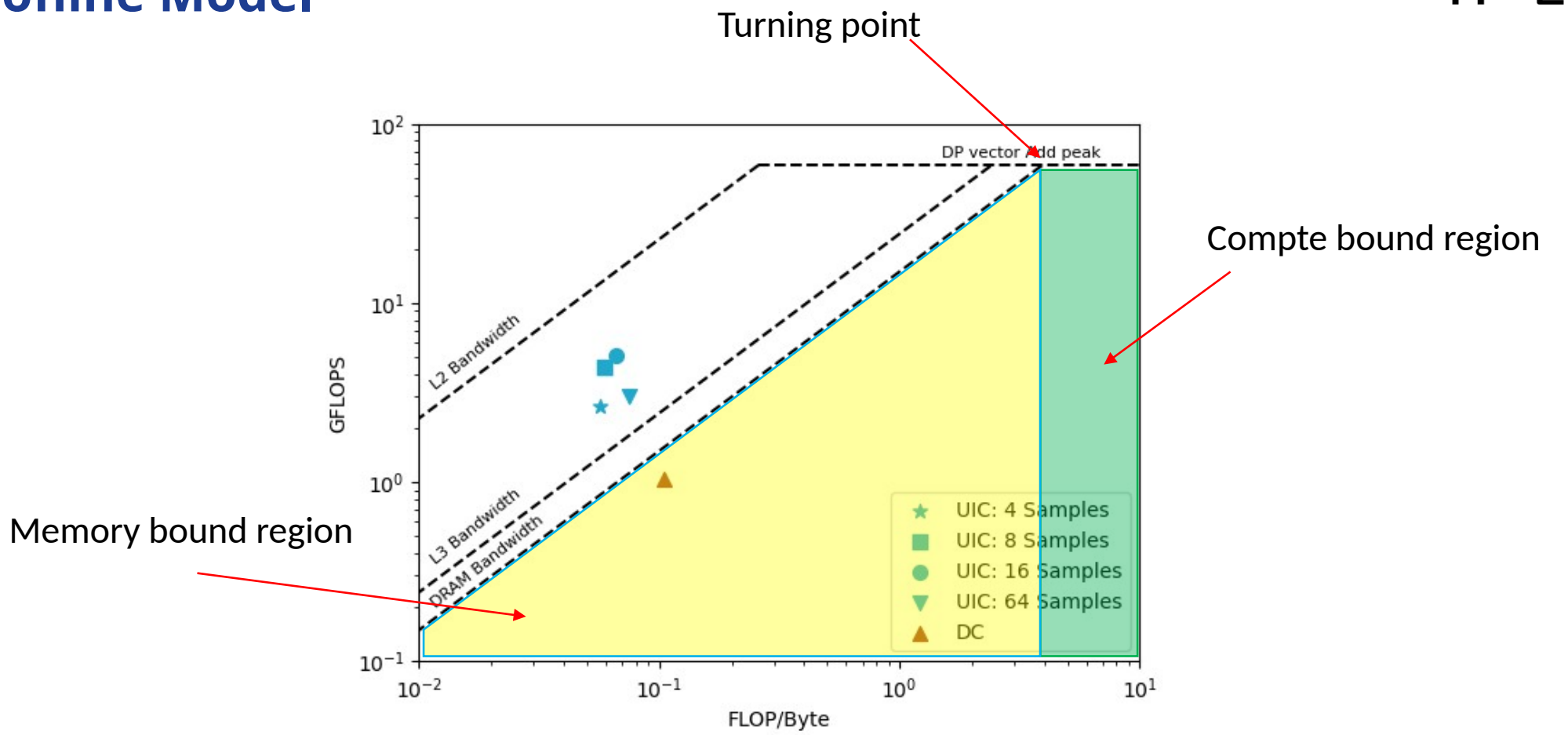
https://huggingface.co/blog/tngtech/llm-performance-prefill-decode-concurrent-requests

LLM Models

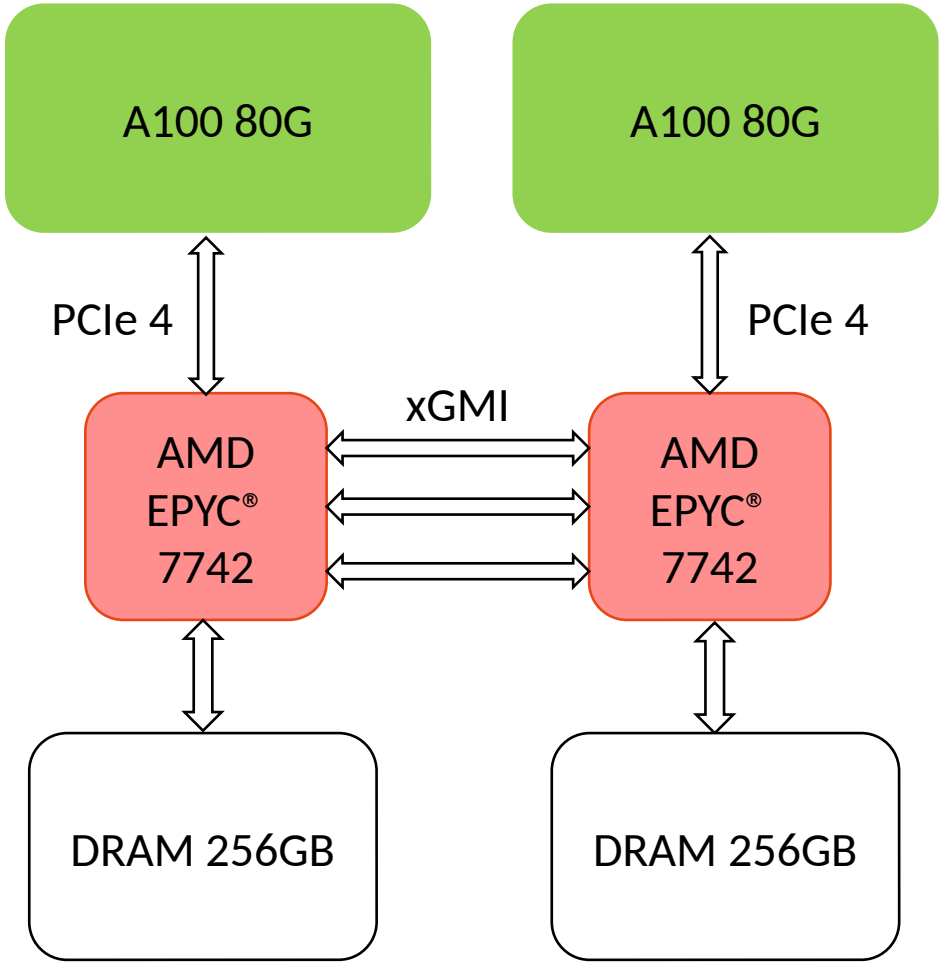


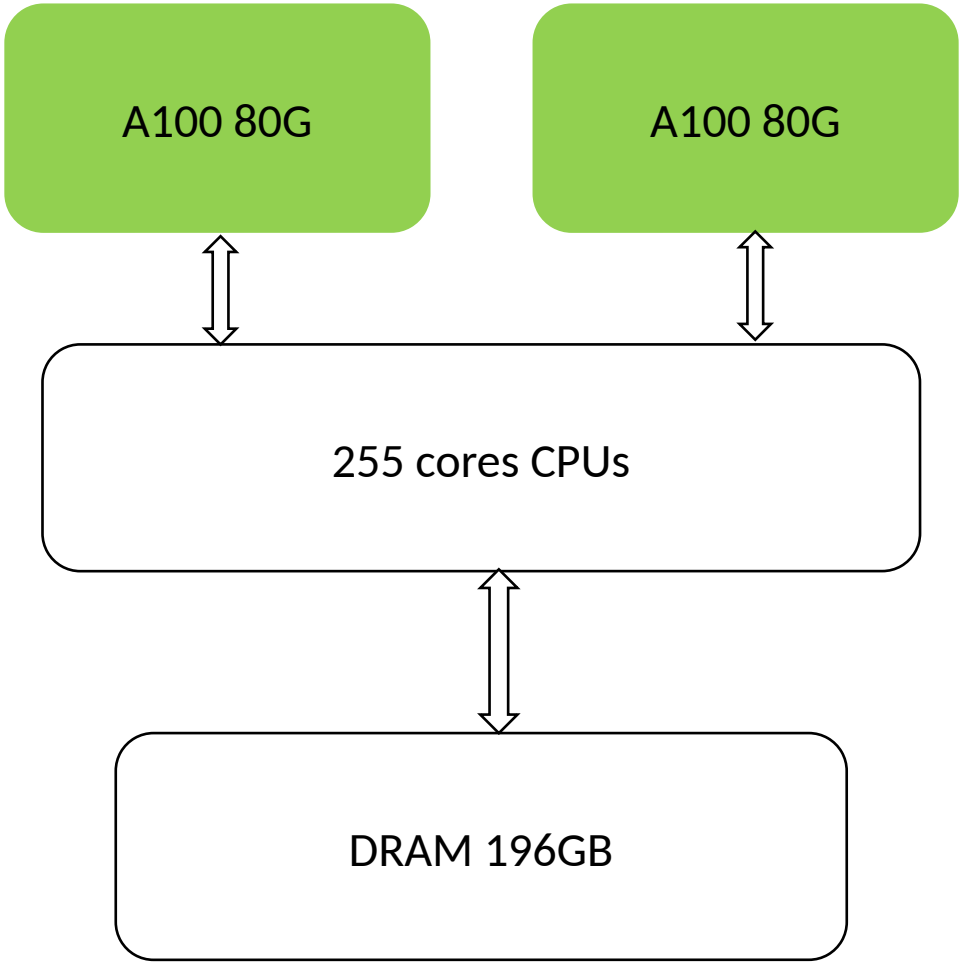
https://huggingface.co/blog/tngtech/llm-performance-prefill-decode-concurrent-requests

Roofline Model

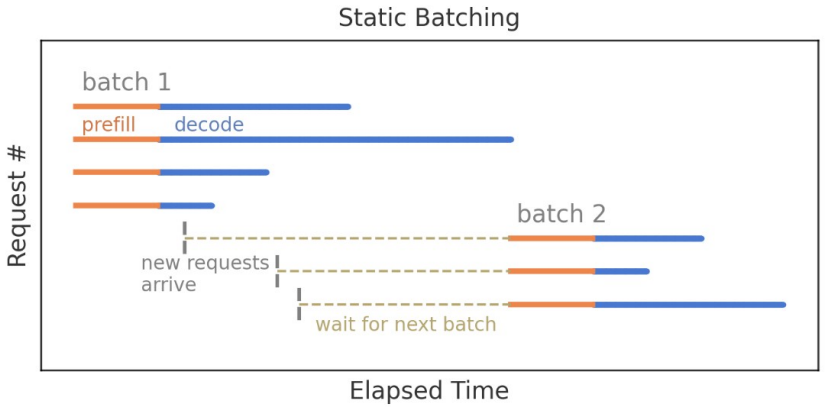


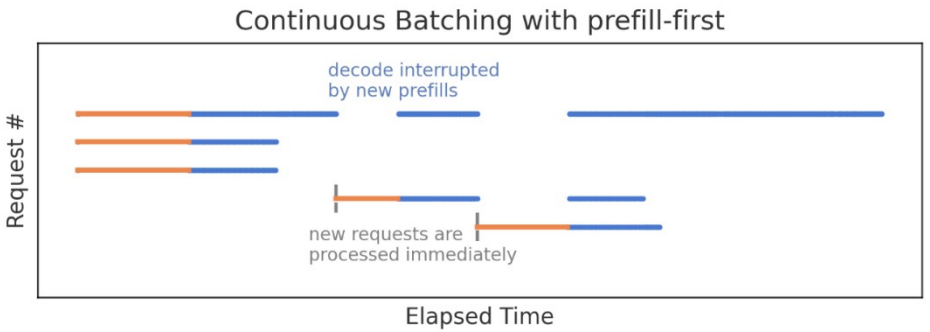
Test environment

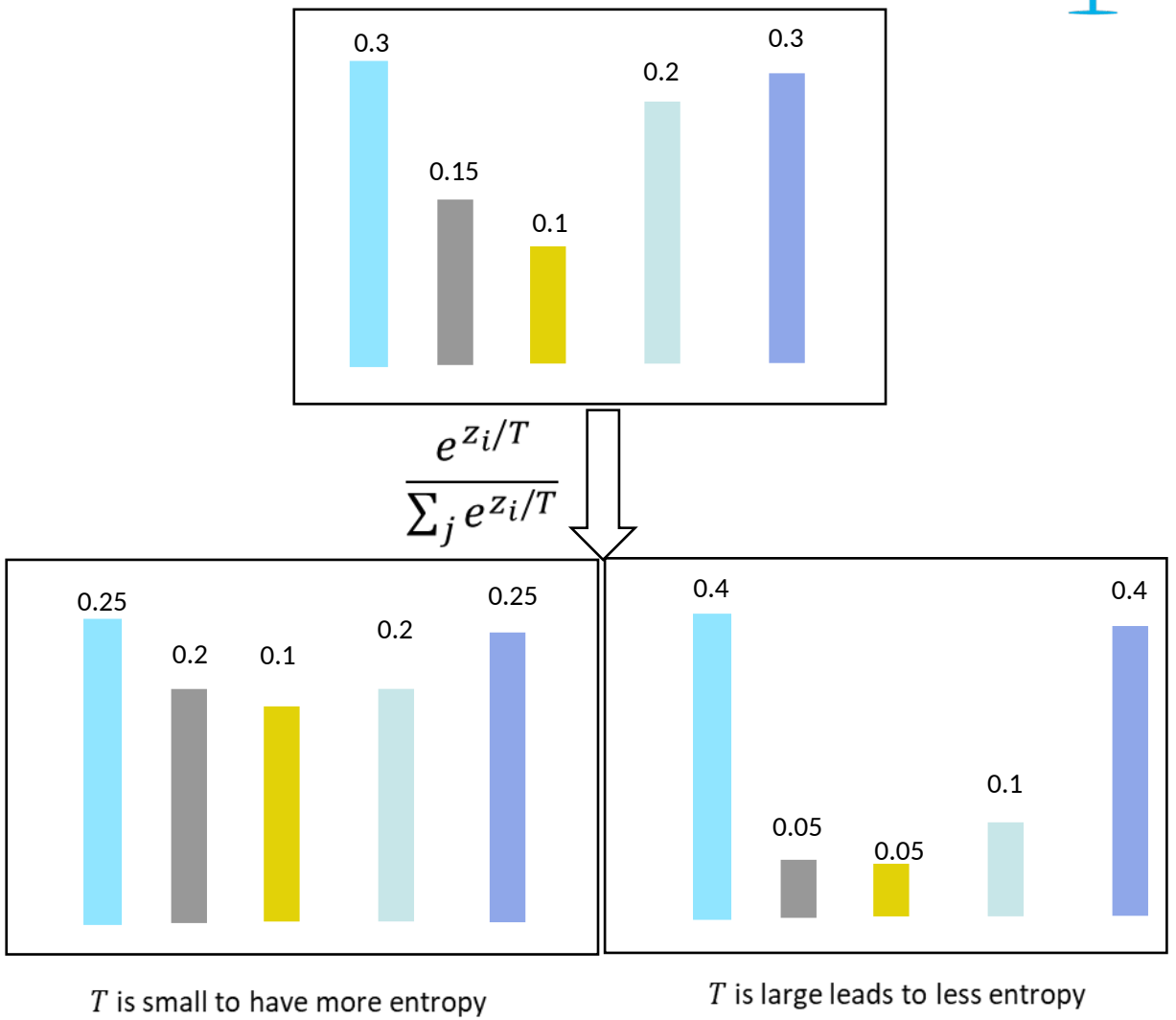




Reproducibility







https://huggingface.co/blog/tngtech/llm-performance-prefill-decode-concurrent-requests

Reproducibility

- Scheduling: Still using the asynchronized configure to ensure the production environment
- Dataset: shareGPT dataset, with seed and no shuffling
- Model parameters:

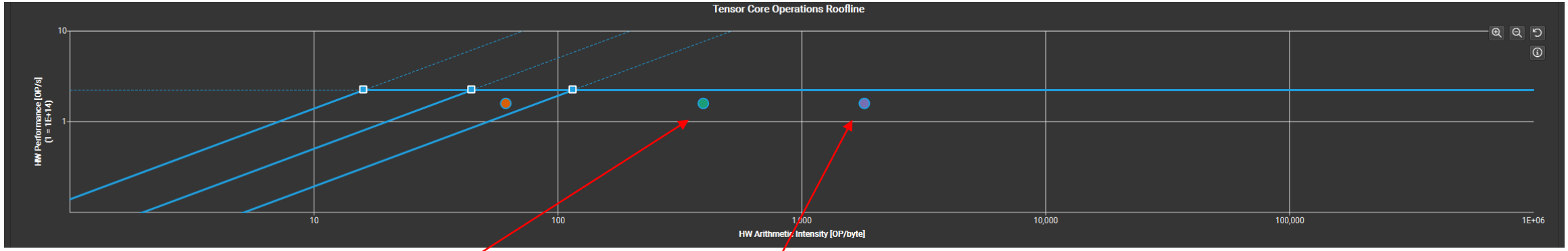
  For production: 0.1 to avoid generate some gibberish

  For profiling: 0 to ensure reproducibility

Preliminary Results

Synthetic pytorch workload:

BF 16 Matrix Matrix multiplication and Matrix Vector multiplication



VM: 399.9 ± 1.26
BM: 404.6 ± 0.629

VM: 1817 ± 0.96
BM: 1818 ± 0.87

VM: $157.45 TFLOPS \pm 0.362$
BM: $158.02\text{TFLOP} \pm 0.309$

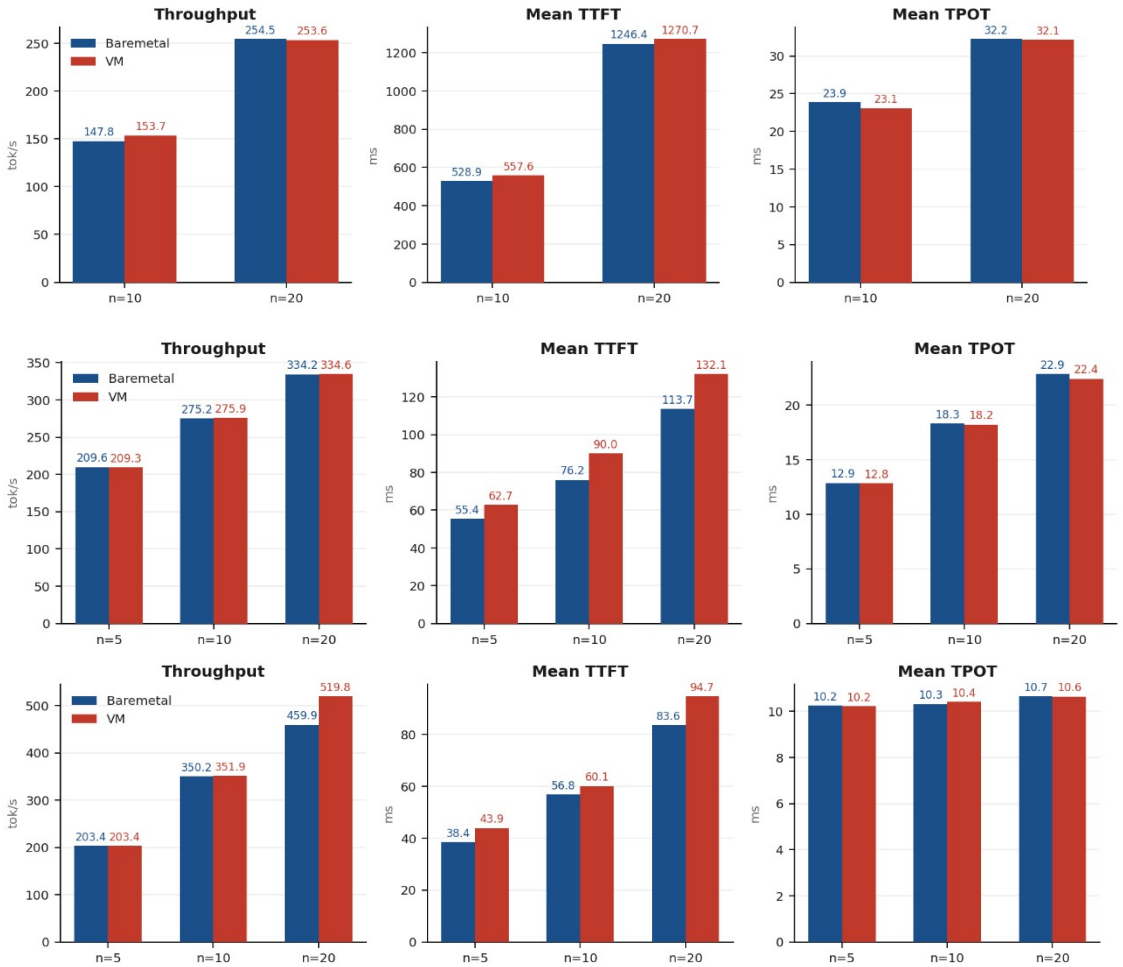
Preliminary Results

One GPU with different workflows

Qwen 3.5-27B
Performance penalty on TTFT:
5.4% and 1.9% for N=10 and 20

Qwen 3.5-30B A3B(MoE model):
Performance penalty on TTFT:
14.1%, 18.2% and 16.2%

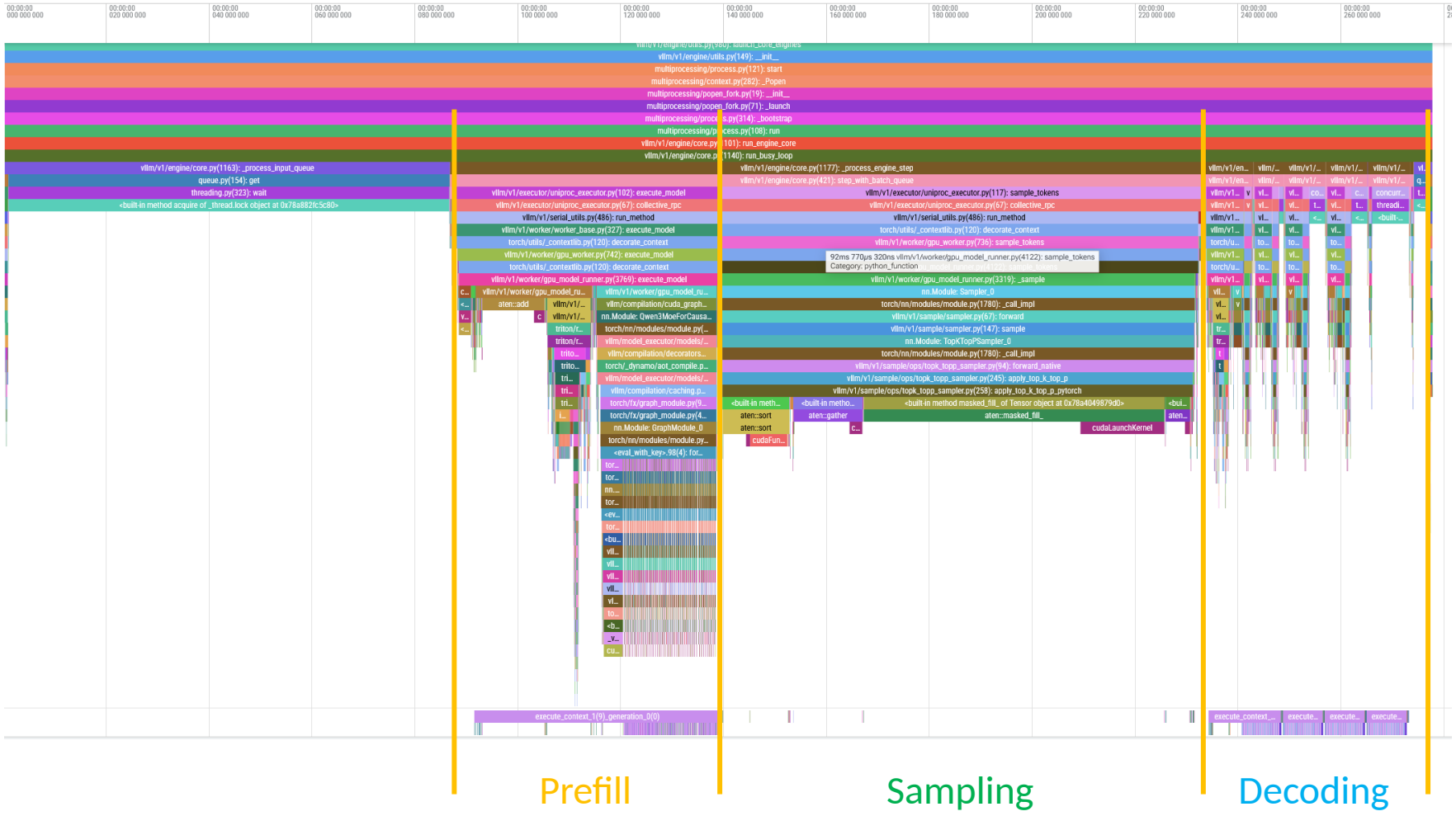
DeepSeek R1 distilled:
Performance penalty on TTFT:
14.1%, 5.8% and 13.3%



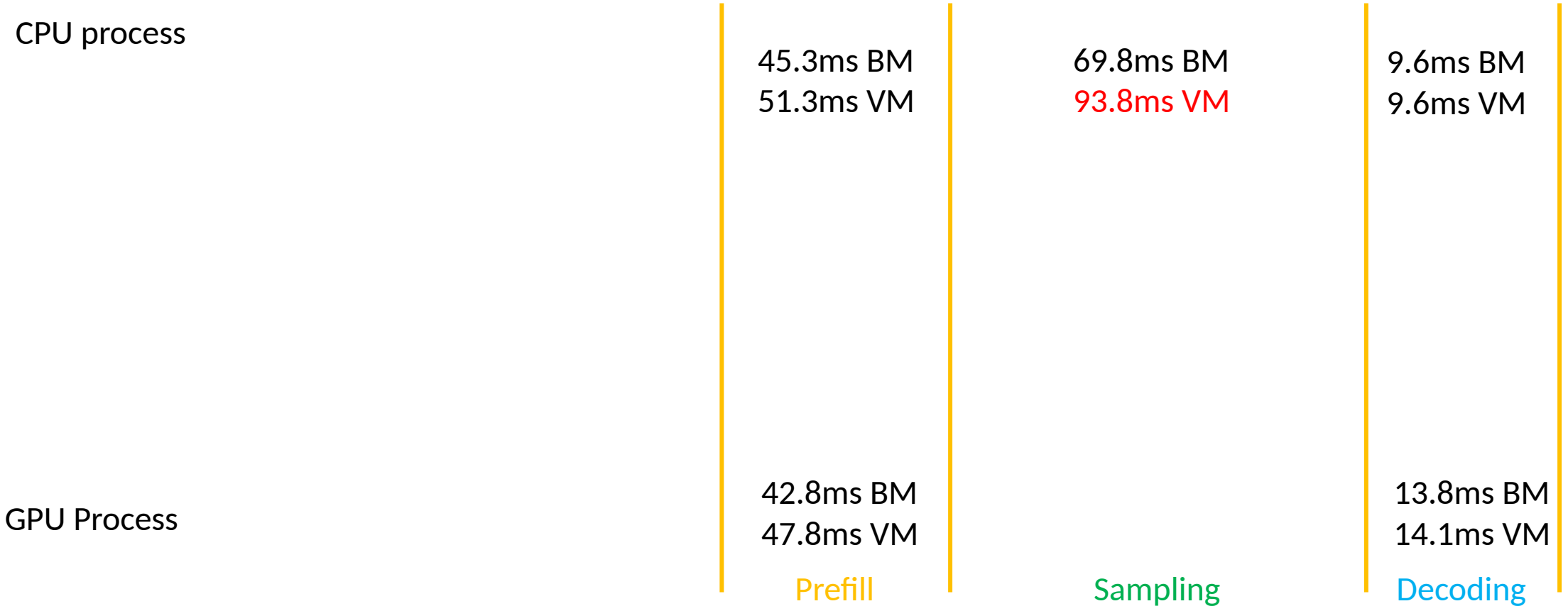
Preliminary Results: Profile for short prompt

CPU process

GPU Process



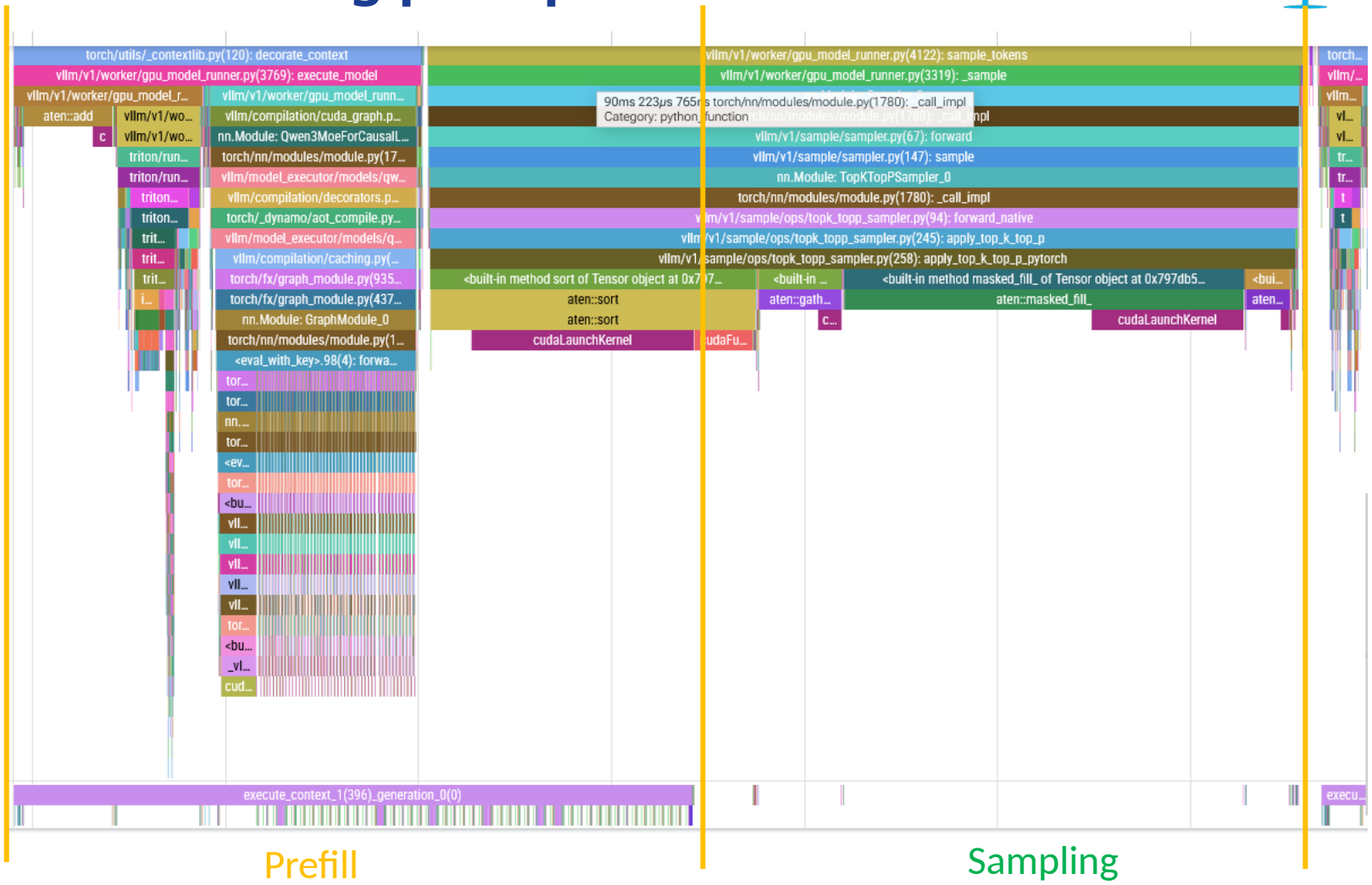
Preliminary Results: Profile for short prompt



Preliminary Results: Profile for long prompt

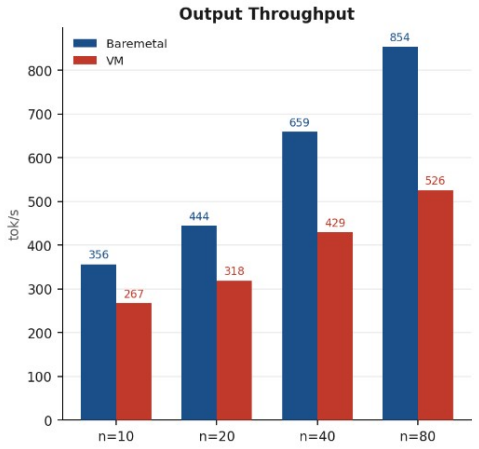
CPU process

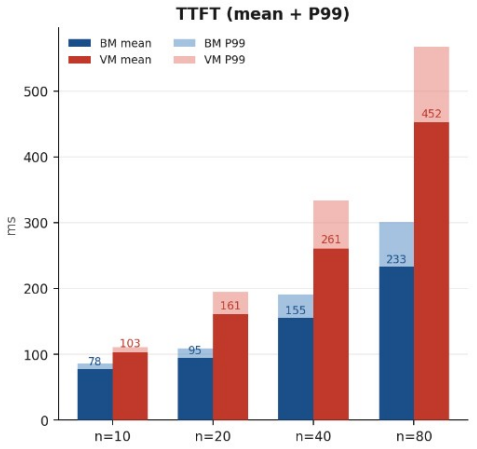
GPU Process

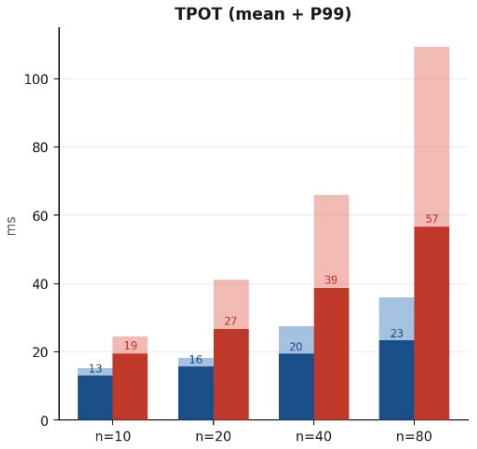


Preliminary Results: Test with Two GPUs

Model Used: Qwen 30B A3B(MoE)
Prompt size: 10,20,40,80

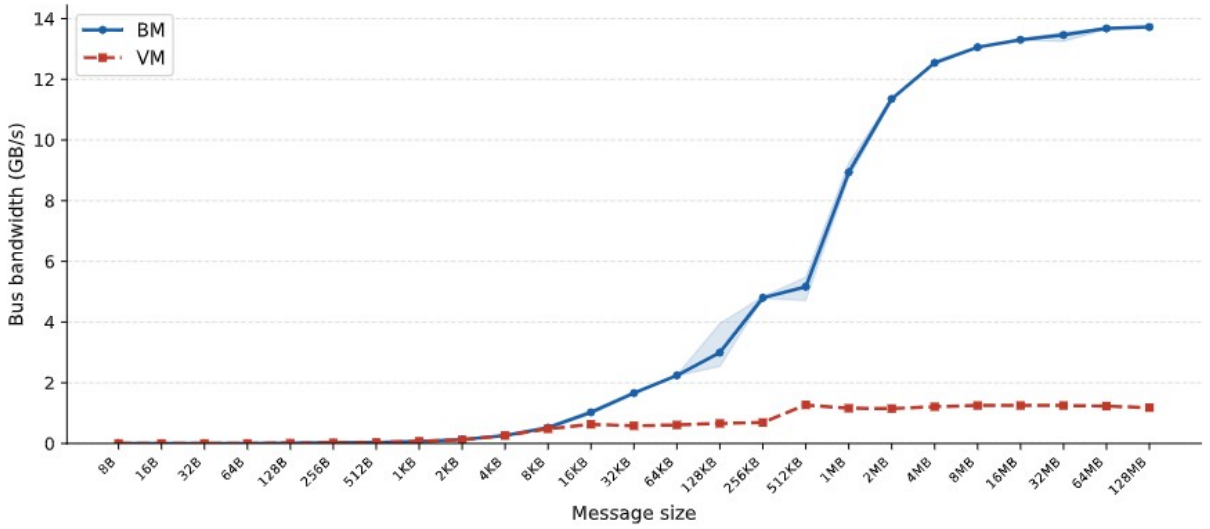


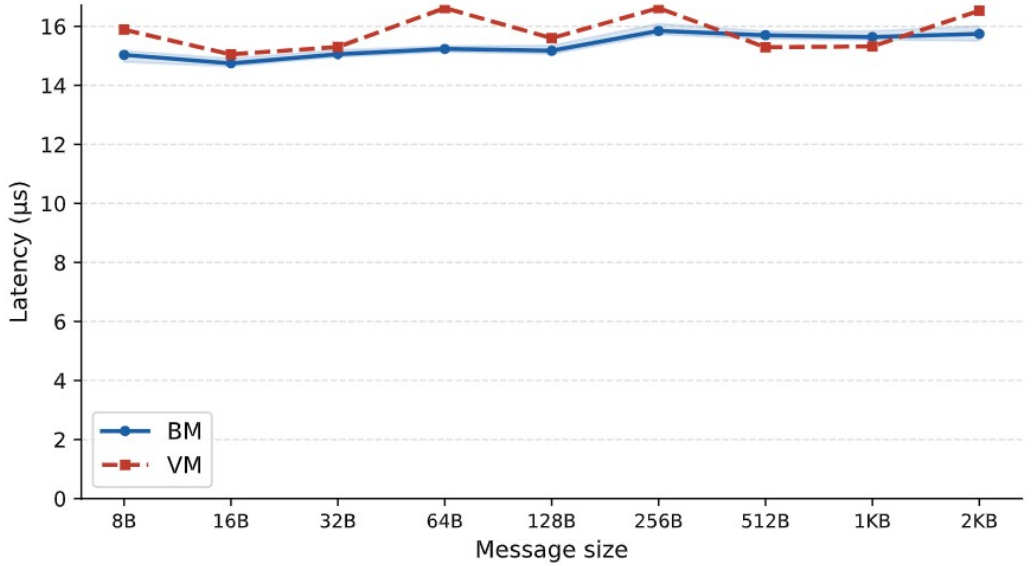




Preliminary Results

The culprit: P2P communication is missing in VM
NCCL test shows a comparable latency but extreme low bandwidth:
BM 14GB/s vs VM 1.2GB/s for All_reduce





Summary and Furture works

- Take aways:

  For small size inferencing fit to 1 GPU, Small performance penalty on Prefill stage
  Barely see performance penalty on decoding stage

  For Large scale model, memory copy and communication can be a headache

- Working on memory copy and NCCL issue, network issue
- DP schemes needs to be added
- Tests on other frameworks such as DeepSpeed
- More in depth profiling and Kernel analysing with roofline model