UNIVERSITY OF CAMBRIDGE

Topic: vLLM Multi-Objective Bayesian Optimization

Presenter: Woon Yee

UNIVERSITY OF CAMBRIDGE

What’s vLLM? What’s Paged Attention?

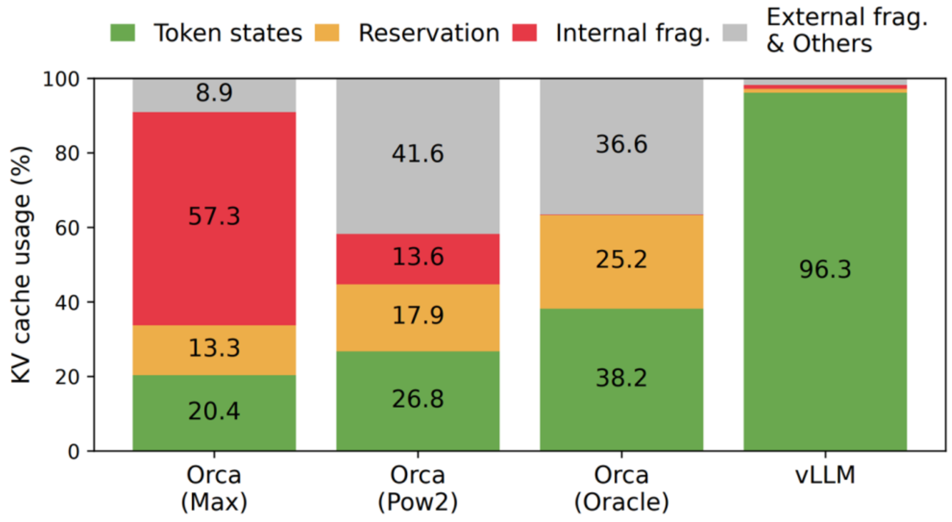


Figure 2. Average percentage of memory wastes in different LLM serving systems during the experiment in §6.2.

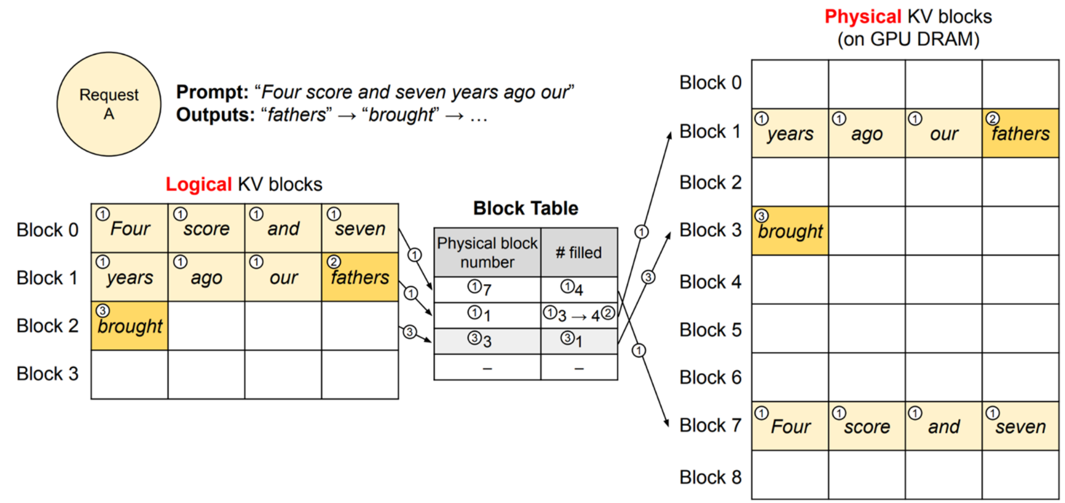


Figure 6. Block table translation in vLLM.

UNIVERSITY OF CAMBRIDGE

What’s vLLM? What’s Paged Attention?

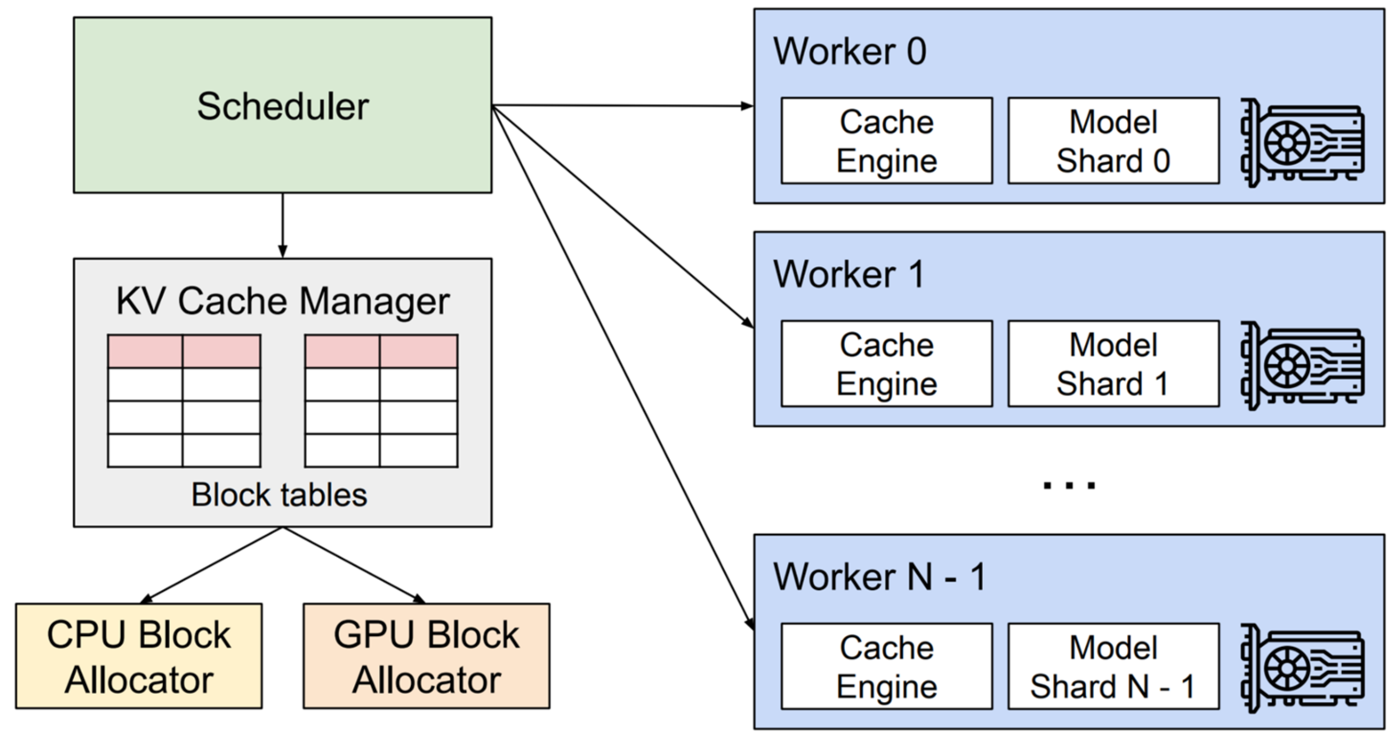


Figure 4. vLLM system overview.

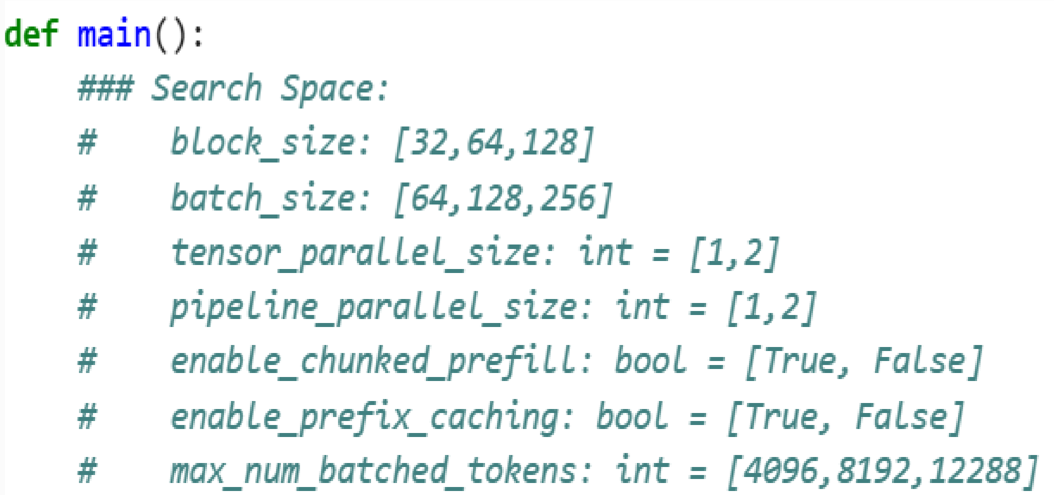
Pain Point:
vLLM configuration is static across workloads!

Goals

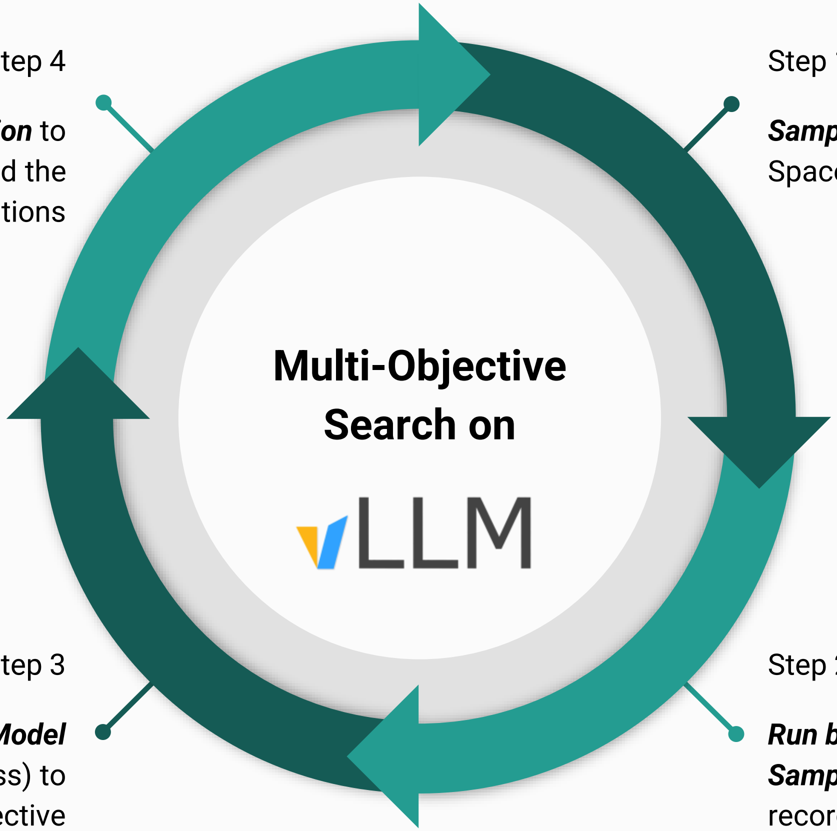
Search for the best vLLM configuration that **minimize energy consumption and maximize throughput**, showed by Pareto Frontier.

UNIVERSITY OF CAMBRIDGE

Multi-Objective Search on vLLM Configurations



```
def main():
    ### Search Space:
    #    block_size: [32,64,128]
    #    batch_size: [64,128,256]
    #    tensor_parallel_size: int = [1,2]
    #    pipeline_parallel_size: int = [1,2]
    #    enable_chunked_prefill: bool = [True, False]
    #    enable_prefix_caching: bool = [True, False]
    #    max_num_batched_tokens: int = [4096,8192,12288]
```



Multi-Objective Search on vLLM

Step 1

Sampling from the Search Space

Step 2

Run benchmark on the Sampled configuration and record the throughput and energy consumption

Step 3

Set Up ***Surrogate Model*** (Gaussian Process) to predict each objective

Step 4

Use ***Acquisition Function*** to assess configs and find the next best configurations

Benchmark for Throughput and Energy

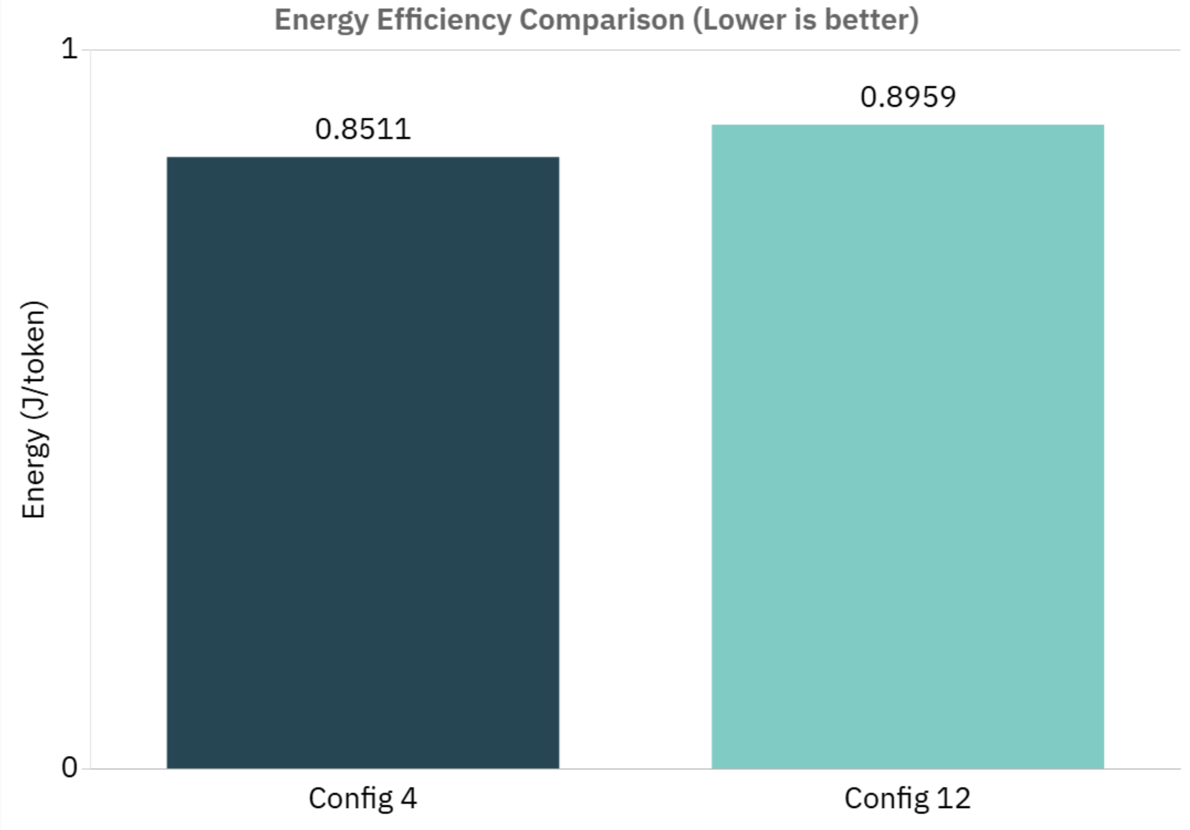
Throughput

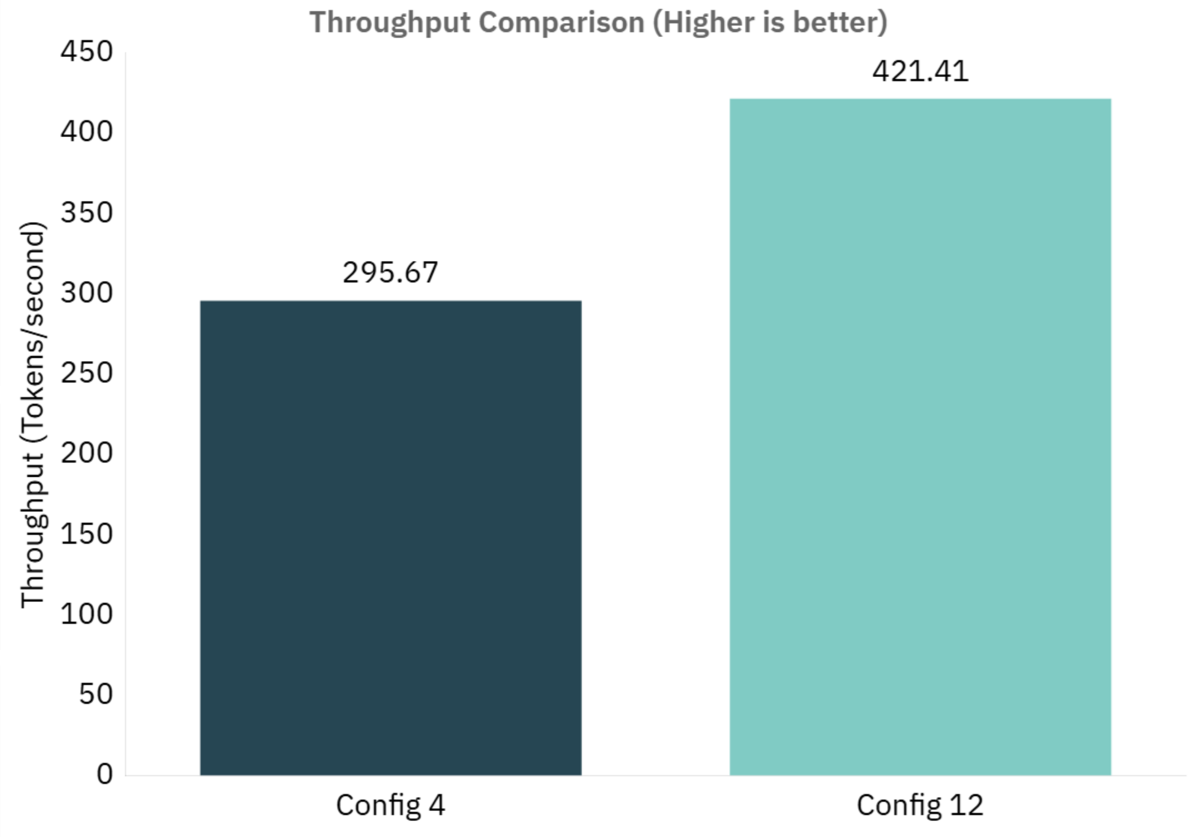
$$\frac{\text{Tokens}}{\text{second}}$$

Energy

1. CPU Energy (Intel RAPL)
2. GPU Energy (NVIDIA NVML or nvidia-smi)
3. RAM Energy (psutil)

Current Progress (Found 2 Pareto Frontiers for vLLM batch Inference of 1000 samples)





Configuration 4 (Energy Optimized):
{ Block size: 128, Batch Size: 256, **Tensor Parallel: 1**, Pipeline Parallel: 2, **Max Batched Token: 4096**, Chunked Prefill: Disabled, Prefix Caching: Enabled }

Configuration 12 (Throughput Optimized):
{ Block size: 128, Batch Size: 256, **Tensor Parallel: 2**, Pipeline Parallel: 2, **Max Batched Token: 12288**, Chunked Prefill: Disabled, Prefix Caching: Enabled }

To Do List

[/] Code for vLLM Experiment

[/] Code for MultiObjective Loop

[/] Code for Benchmarking (Throughput and Latency)

[   ] More Experiments, Online Inference instead of Batch Inference?

[   ] nsys profiling on the best configuration to understand tp/pp trade off

[   ] Design an **energy-driven** acquisition function in MOBO?

[   ] Result Analysis (Graph Plotting and Reasoning)

[   ] Report Writing

Thank you

Questions?