清華大學
Tsinghua University

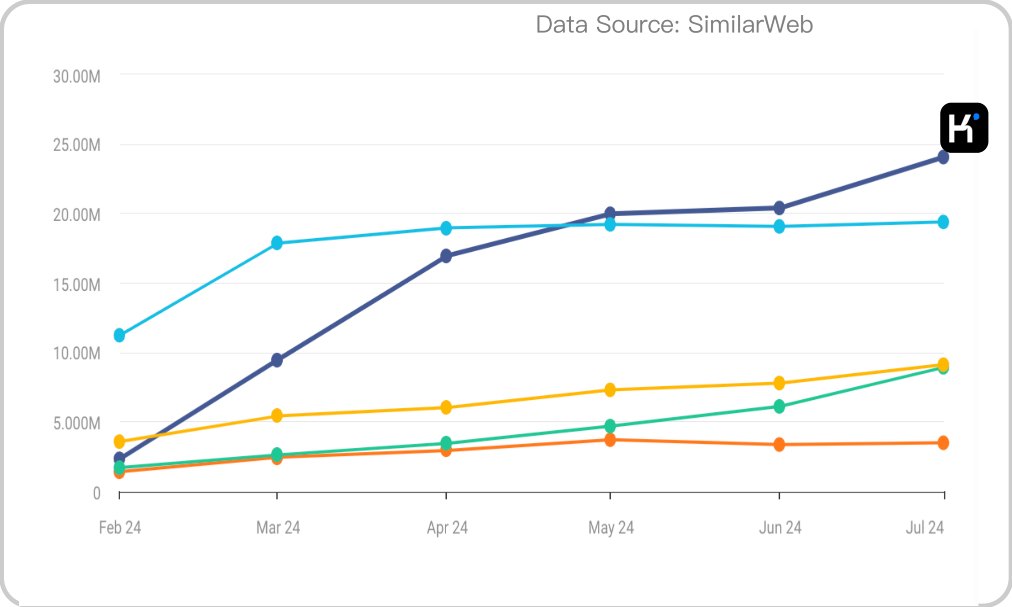
KTransformers: Unleashing the Full Potential of CPU/GPU Hybrid Inference for MoE Models

Mingxing Zhang @ KVCache.AI

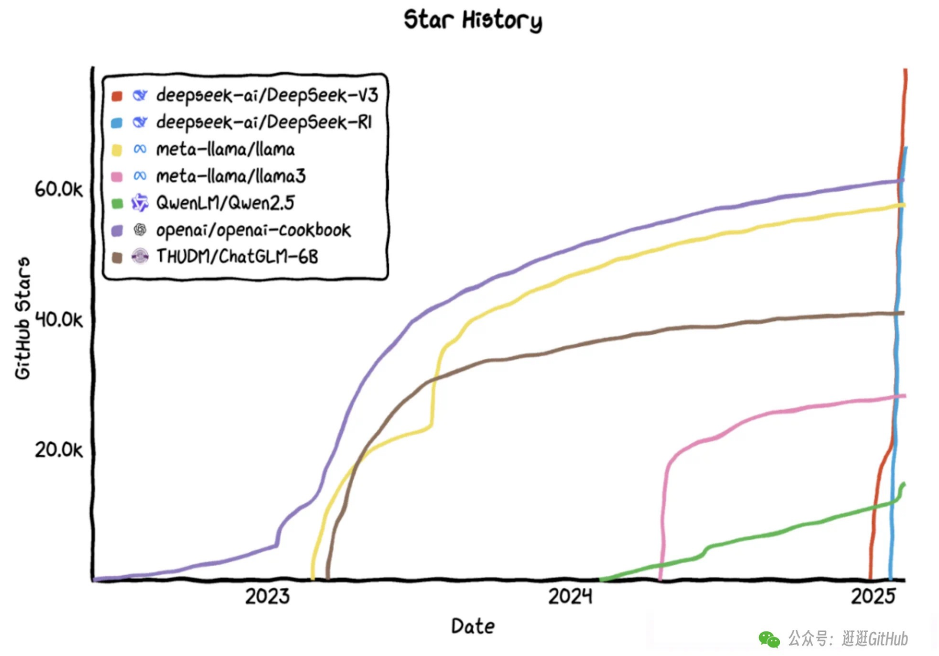
https://github.com/kvcache-ai

Challenge of Online Model as a Service System

More Data + Larger Model + ***Longer Context*** = ☺ Higher Intelligence


Data Source: SimilarWeb

Long input: Moonshot AI's Kimi Supports 2 Million Characters Input in March 2024, become a widely recognized app in China

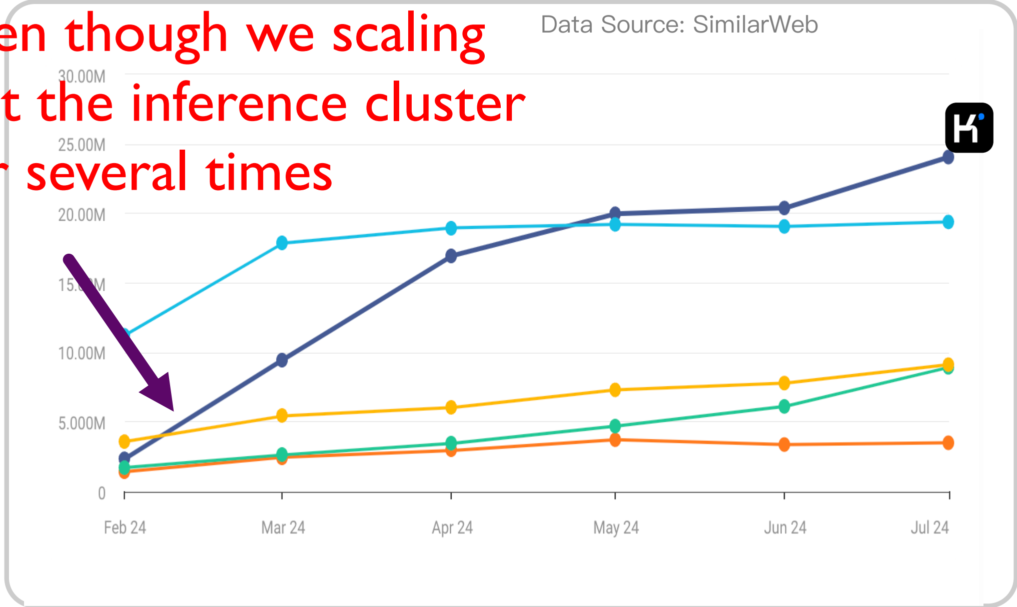


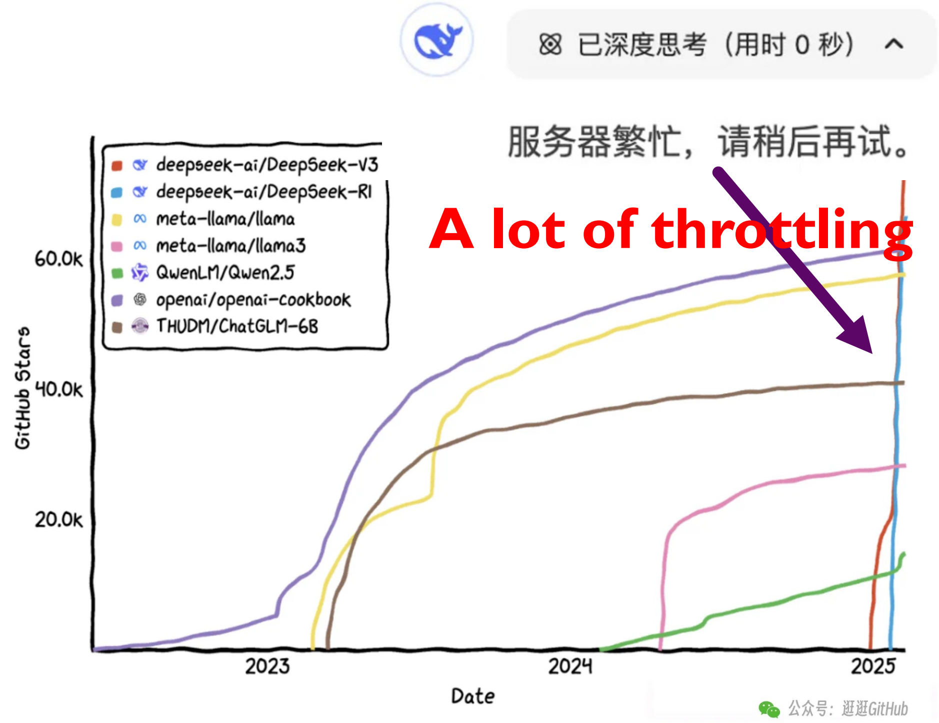
Long output: DeepSeek release V3/R1 at Dec 2024, Become a widely recognized app in global

Challenge of Online Model as a Service System

More Data + Larger Model + ***Longer Context*** = ☹ Higher Service Loads

Frequent out of service even though we scaling out the inference cluster for several times




已深度思考（用时 0 秒）

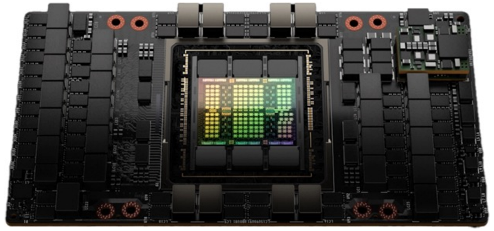
服务器繁忙，请稍后再试。

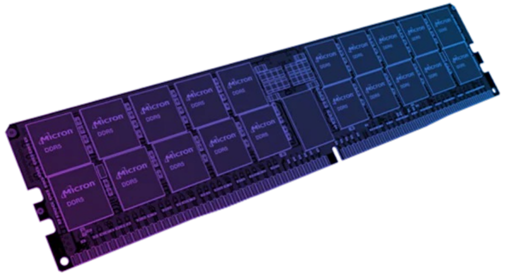
A lot of throttling

Long input: Moonshot AI's Kimi Supports 2 Million Characters Input in March 2024, become a widely recognized app in China

Long output: DeepSeek release V3/R1 at Dec 2024, Become a widely recognized app in global

Different Hardware are Good at Different Dimension



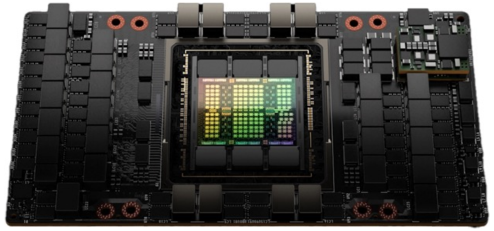


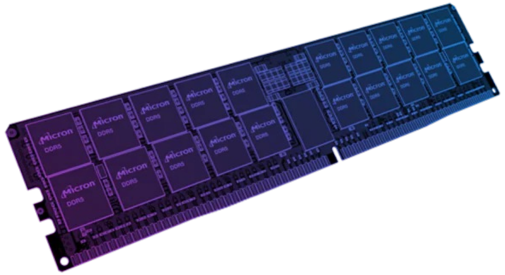
	H800	Xeon SPR + 8 * DDR5-4800
Hardware Spec	80GB VRAM， 3.3 TBps ~ 1 PFLOPS > $ 10,000	8*64GB DRAM， 8*40GB/s < 20 TFLOPS ~ ¥60,000
Best for	Allround, especially for TFLOPS/$	Capacity/$

Good for Capacity, bad for Bandwidth and Compute Which part is more suitable?

!!! The price numbers are not accurate, just a demonstration!

Different Hardware are Good at Different Dimension





	H800	Xeon SPR + 8 * DDR5-4800
Hardware Spec	80GB VRAM， 3.3 TBps ~ 1 PFLOPS > $ 10,000	8*64GB DRAM， 8*40GB/s < 20 TFLOPS ~ ¥60,000
Best for	Allround, especially for TFLOPS/$	Capacity/$

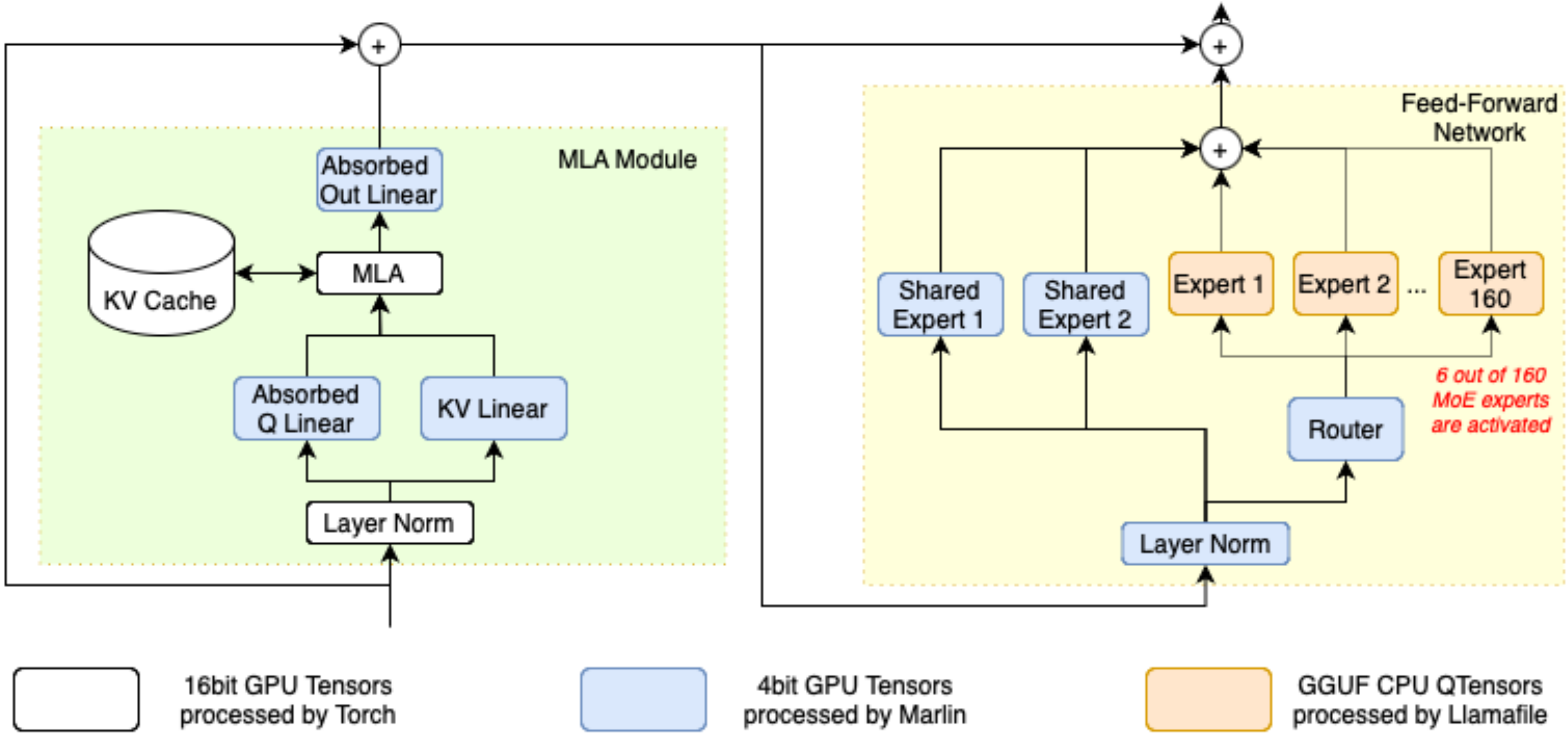
Good for Capacity, bad for Bandwidth and Compute
Which part is more suitable?

Sparsity!

!!! The price numbers are not accurate, just a demonstration!

Take DeepSeek as an Example

- DeepSeek Archtecture



Offload Priority

Routed Experts

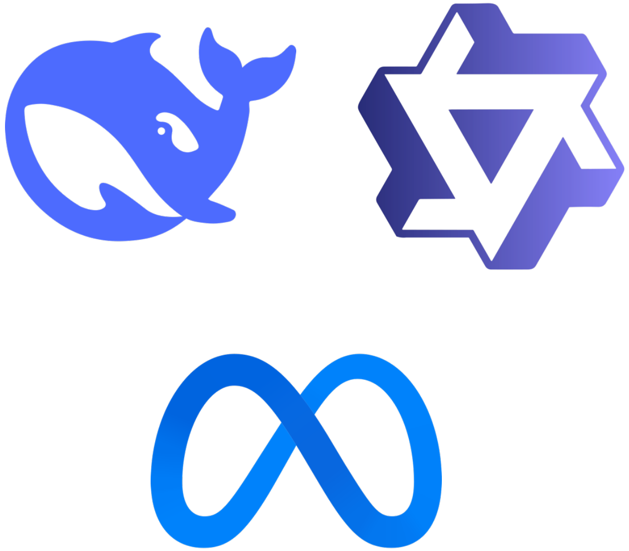
>

Shared Experts

>

MLA Attention

Not Only DeepSeek



Different Models

×

Different Hardware

Challenges

Decode

Latency, again,

the latencies!

CUDAGraph

is the key

(only one launch each forward)

Prefill

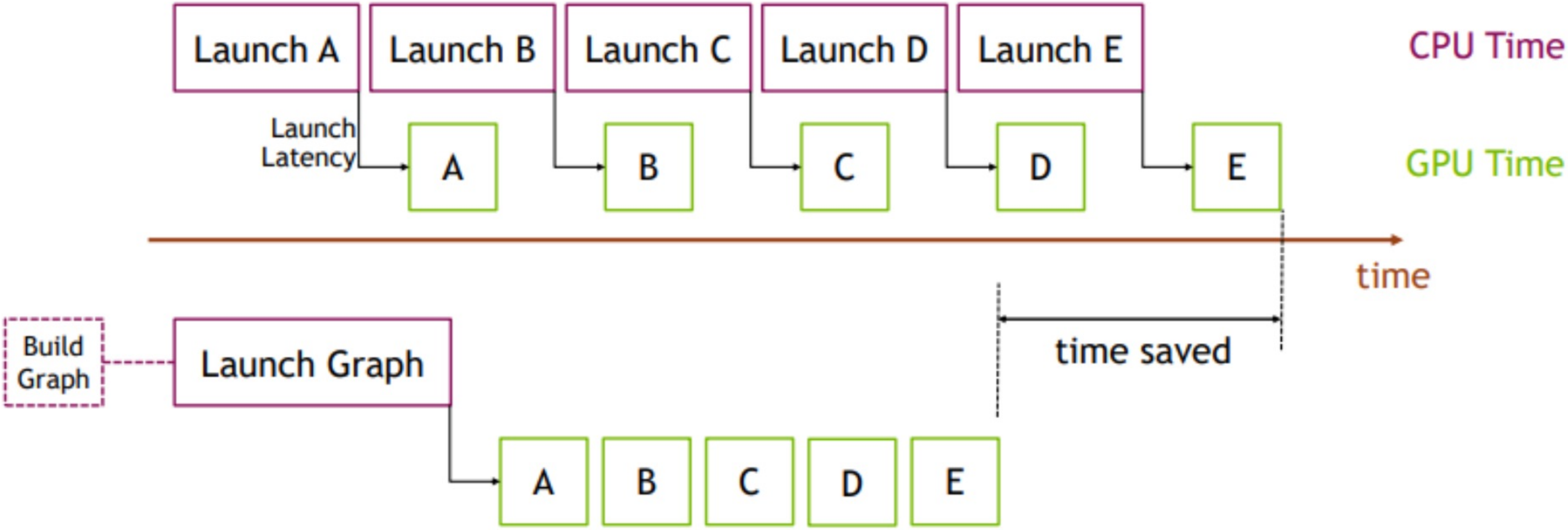
CPU is too weak, even GQA 8

becomes compute bound

New hardware:

Intel AMX

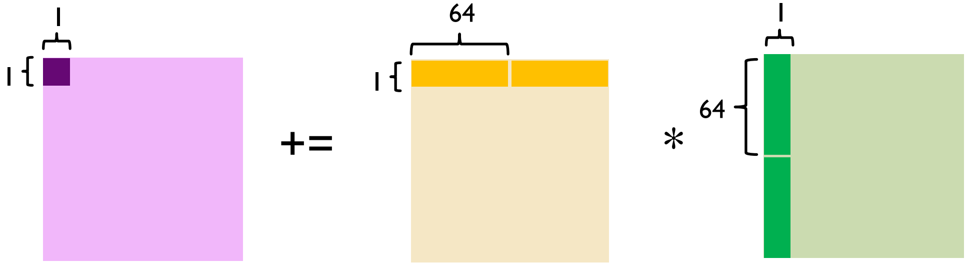
CUDA Graph



How to handle dynamic shape in continues-batched decoding?

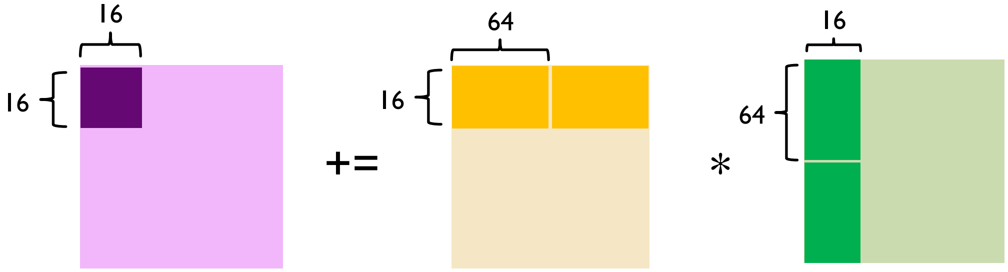
Intel Advanced Matrix Extensions (Intel AMX)

How AVX-512 solves INT8 matrix multiplication problems



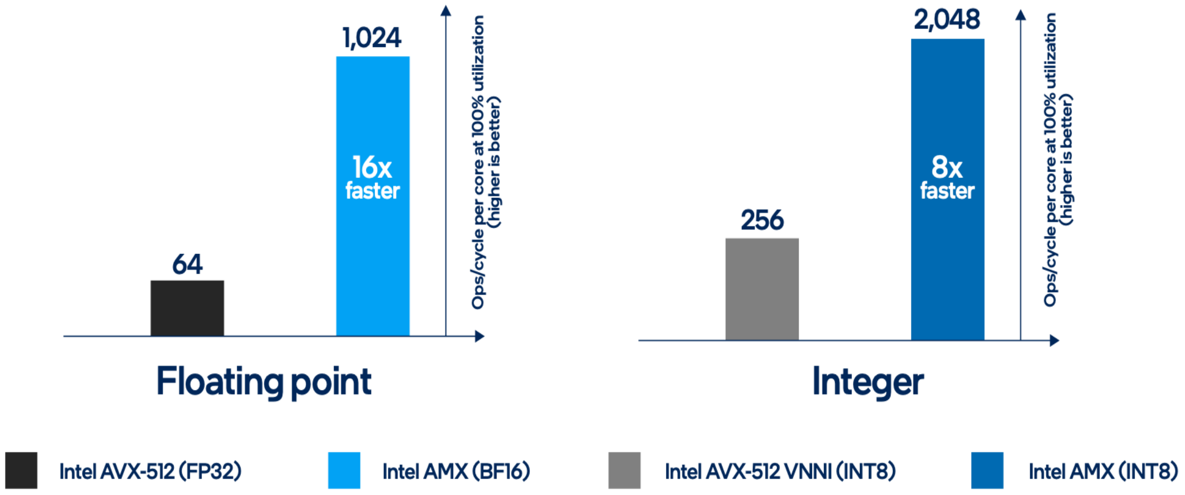
128OPS/cycle/FMA. 256OPS/cycle/core

How AMX solves INT8 matrix multiplication problems



32768OPS/16cycle/core. 2048OPS/cycle/core

AMX offers better performance than AVX-512 for INT8 and BF16 data types.



Key optimization of matrix multiplication with Intel AMX

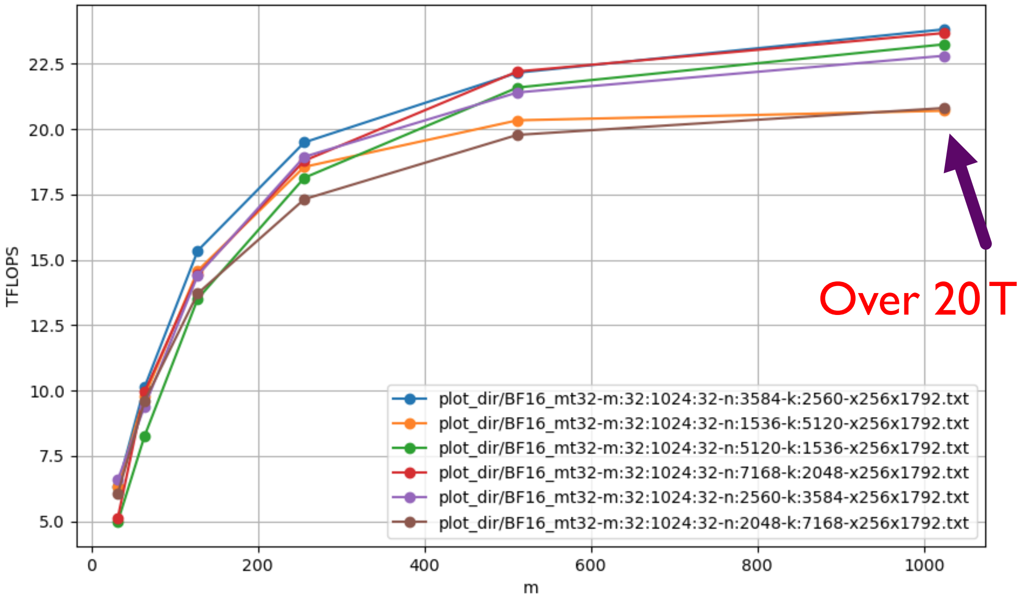
Due to the high computational capability of AMX instructions, memory bandwidth becomes a bottleneck, and the key to optimization lies in improving cache hit rates.

1. Design memory layout based on access patterns to enhance data locality, fully leverage hardware prefetch capabilities, and improve **L1 cache** hit rates.
2. Split large matrices into smaller tiles based on L2 cache size, ensuring that only the current tile is accessed at a time, improving **L2 cache** hit rates.
3. Dynamic work-scheduling， increasing data sharing between threads and optimizing **L3 cache** hit rates.
4. Quantize matrices by rows/columns, maintaining precision while reducing the number of scaling operations (and memory access) for each output element.

Applications of Intel AMX

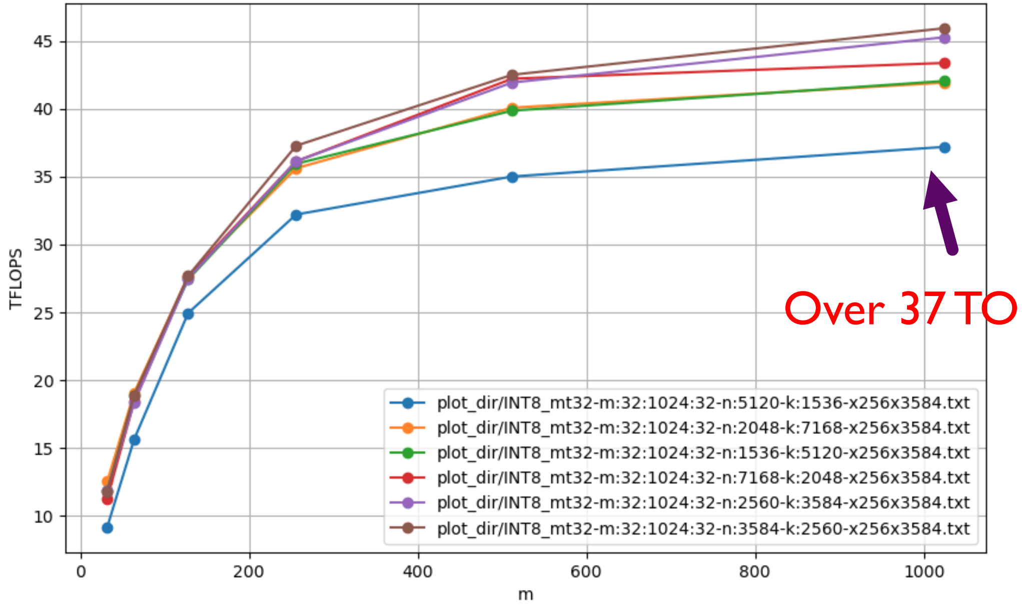
Applications： AMX kernel for sparse layers of MoE models (Deepseek R1/V3/V2, Mixtral, etc.)

BF16 TFLOPS vs. avg. # of selections



Over 20 TFLOPS

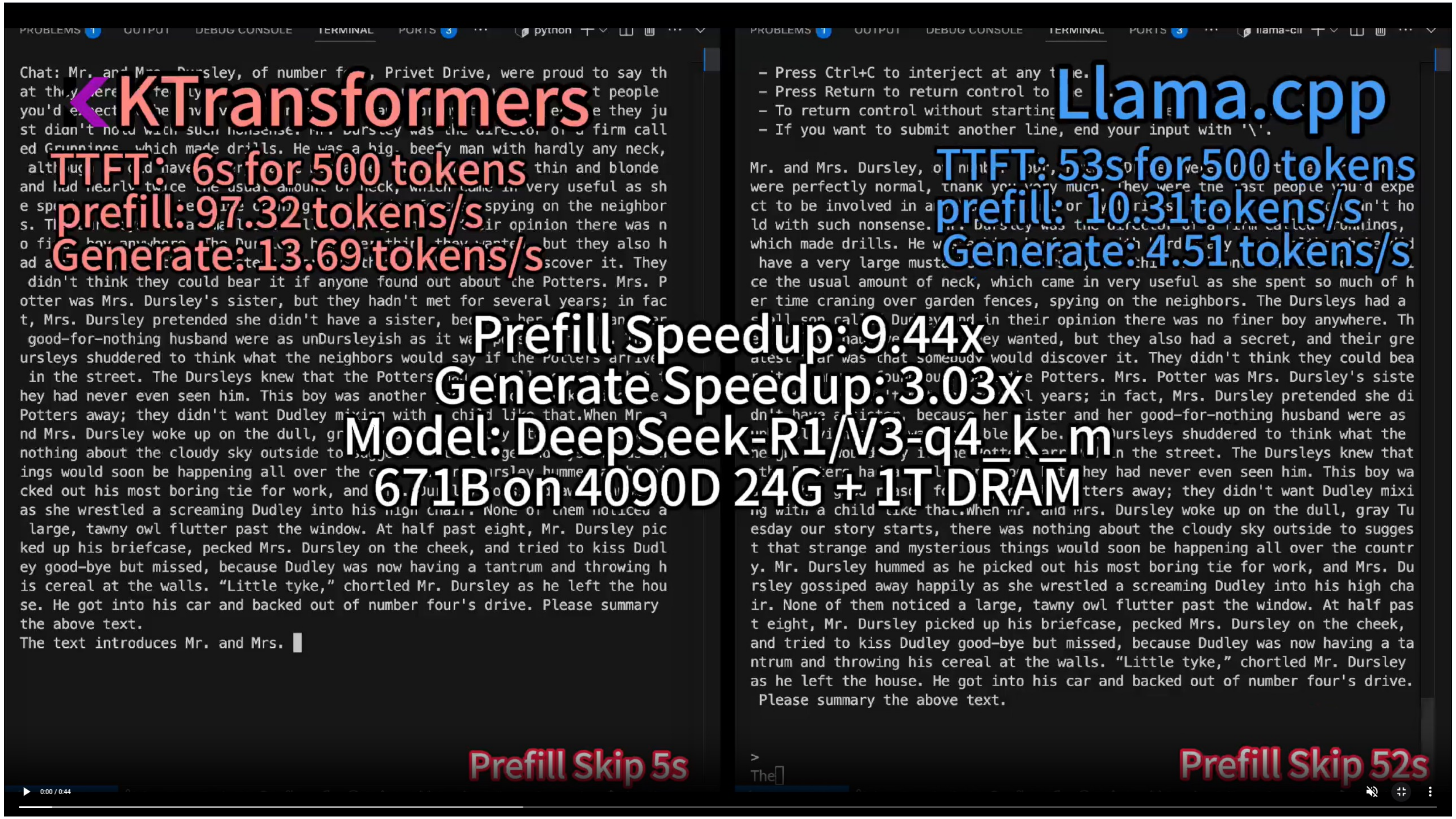
INT8 TOPS vs. avg. # of selections



Over 37 TOPS

In the matrix multiplication micro-benchmark, achieve over 20 TFLOPS and 37 TOPS of computational performance.
In the sparse MOE layer, achieve over 18TFLOPS and 30 TOPS of end-to-end computational performance.

The Results: KTransformers



KTransformers

TTFT: 6s for 500 tokens
prefill:97.32 tokens/s
Generate:13.69 tokens/s

```
Chat: Mr. and Mrs. Dursley, of number f[illegible], Privet Drive, were proud to say th
at th[illegible]t people
you'd e[illegible]e they ju
st didn't hold with such nonsense. Mr. Dursley was the director of a firm call
ed Grunnings, which made drills. He was a big, beefy man with hardly any neck,
 alth[illegible] thin and blonde
and had nearly twice the usual amount of neck, which came in very useful as sh
e sp[illegible] spying on the neighbor
s. T[illegible]ir opinion there was n
o fi[illegible] but they also h
ad a[illegible]scover it. They
 didn't think they could bear it if anyone found out about the Potters. Mrs. P
otter was Mrs. Dursley's sister, but they hadn't met for several years; in fac
t, Mrs. Dursley pretended she didn't have a sister, bec[illegible]
 good-for-nothing husband were as unDursleyish as it wa[illegible]
ursleys shuddered to think what the neighbors would say if the Potters arrive
 in the street. The Dursleys knew that the Potters [illegible]
hey had never even seen him. This boy was another [illegible]
Potters away; they didn't want Dudley mixing with [illegible] child like that.When Mr. a
nd Mrs. Dursley woke up on the dull, gr[illegible]
nothing about the cloudy sky outside to [illegible]
ings would soon be happening all over the c[illegible]
cked out his most boring tie for work, and [illegible]
as she wrestled a screaming Dudley into his high chair. None of them noticed a
 large, tawny owl flutter past the window. At half past eight, Mr. Dursley pic
ked up his briefcase, pecked Mrs. Dursley on the cheek, and tried to kiss Dudl
ey good-bye but missed, because Dudley was now having a tantrum and throwing h
is cereal at the walls. "Little tyke," chortled Mr. Dursley as he left the hou
se. He got into his car and backed out of number four's drive. Please summary
the above text.
The text introduces Mr. and Mrs. █
```

Prefill Skip 5s

Llama.cpp

TTFT: 53s for 500 tokens
prefill: 10.31tokens/s
Generate: 4.51 tokens/s

```
 - Press Ctrl+C to interject at any t[illegible]e.
 - Press Return to return control to [illegible]e
 - To return control without starting[illegible]
 - If you want to submit another line, end your input with '\'.

Mr. and Mrs. Dursley, o[illegible]
were perfectly normal, thank you very much. They were the last people you'd expe
ct to be involved in a[illegible]n't ho
ld with such nonsense. [illegible]Grunnings,
which made drills. He w[illegible]
 have a very large musta[illegible]
ce the usual amount of neck, which came in very useful as she spent so much of h
er time craning over garden fences, spying on the neighbors. The Dursleys had a
 [illegible] Dudley and in their opinion there was no finer boy anywhere. Th
e[illegible]ey wanted, but they also had a secret, and their gre
atest [illegible]ar was that somebody would discover it. They didn't think they could bea
[illegible]e Potters. Mrs. Potter was Mrs. Dursley's siste
[illegible]l years; in fact, Mrs. Dursley pretended she di
dn't have a sister, because her sister and her good-for-nothing husband were as
[illegible]ble be. Dursleys shuddered to think what the
[illegible]n the street. The Dursleys knew that
[illegible]hey had never even seen him. This boy wa
[illegible]tters away; they didn't want Dudley mixi
ng with a child like that.When Mr. and Mrs. Dursley woke up on the dull, gray Tu
esday our story starts, there was nothing about the cloudy sky outside to sugges
t that strange and mysterious things would soon be happening all over the countr
y. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Du
rsley gossiped away happily as she wrestled a screaming Dudley into his high cha
ir. None of them noticed a large, tawny owl flutter past the window. At half pas
t eight, Mr. Dursley picked up his briefcase, pecked Mrs. Dursley on the cheek,
and tried to kiss Dudley good-bye but missed, because Dudley was now having a ta
ntrum and throwing his cereal at the walls. "Little tyke," chortled Mr. Dursley
as he left the house. He got into his car and backed out of number four's drive.
 Please summary the above text.

>
The▯
```

Prefill Skip 52s

Prefill Speedup: 9.44x
Generate Speedup: 3.03x
Model: DeepSeek-R1/V3-q4_k_m
671B on 4090D 24G + 1T DRAM

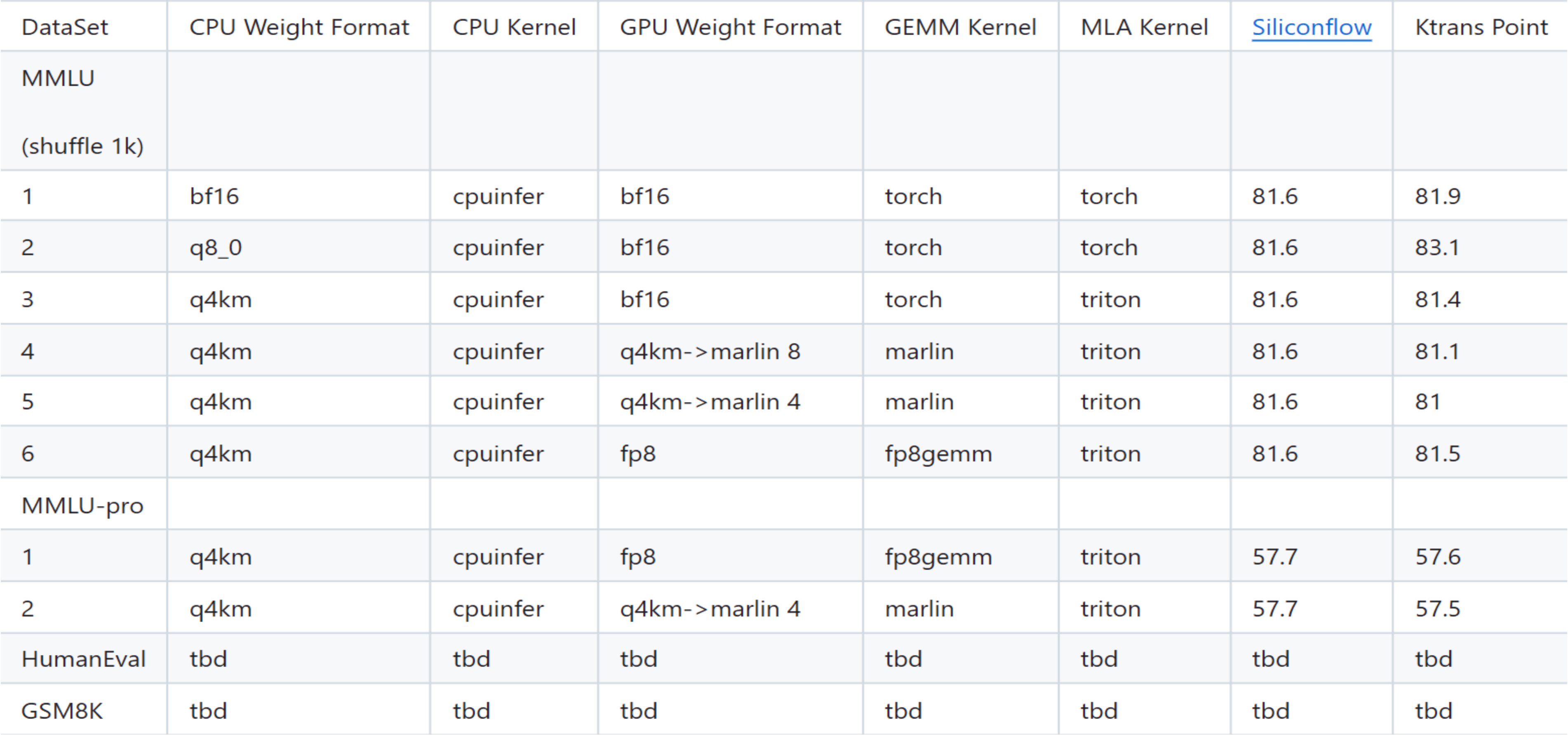
KTransformers

A flexible heterogeneous inference framework

- 24GB VRAM +382GB DRAM for 761B q4
- Currently several times faster than llama.cpp

https://github.com/kvcache-ai/ktransformers

Impact of quantization

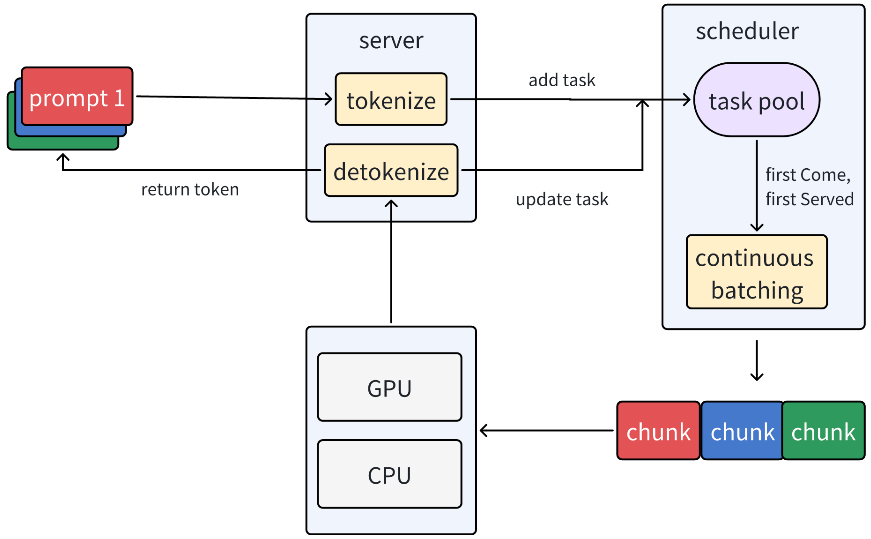


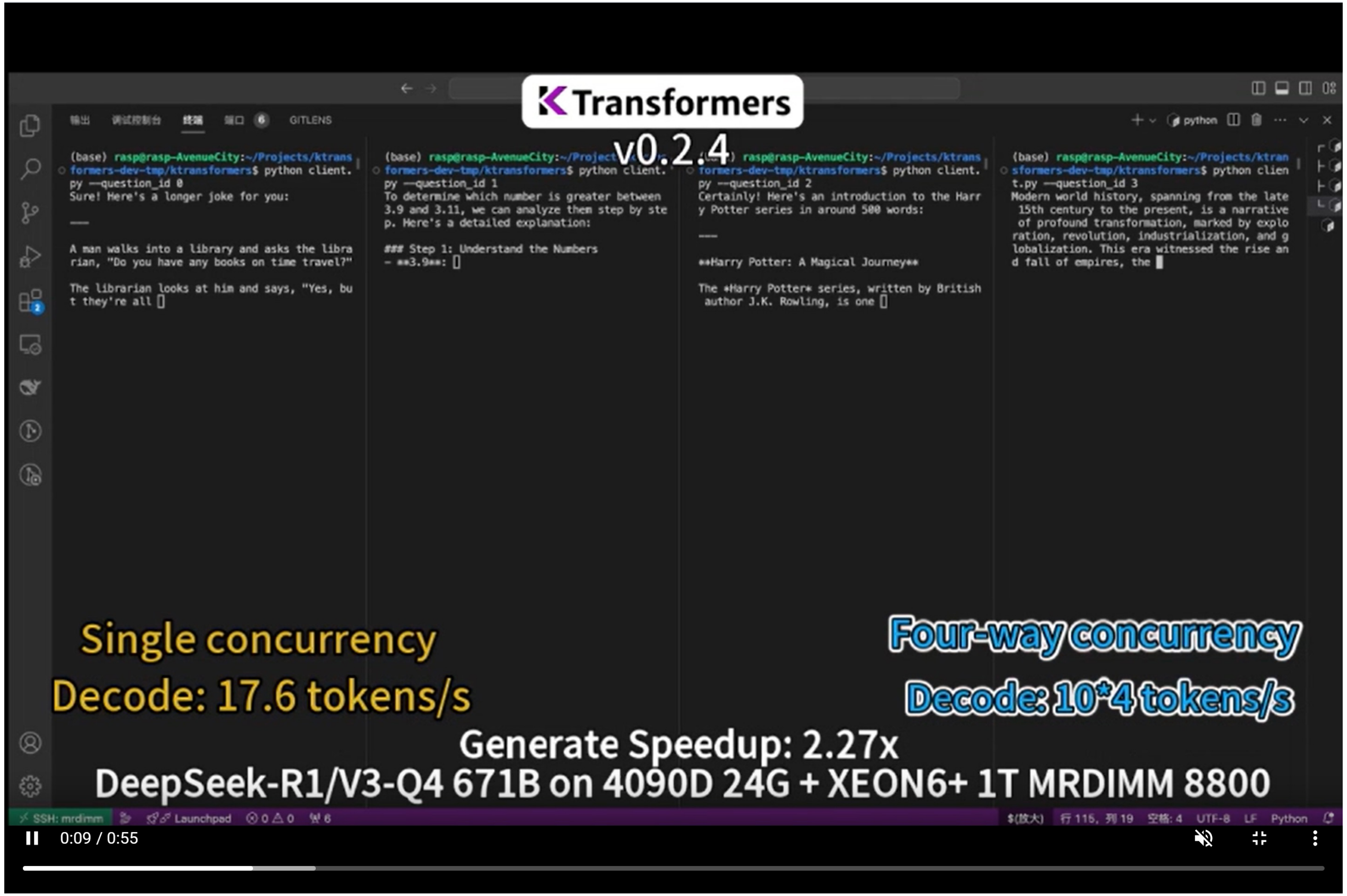
DataSet	CPU Weight Format	CPU Kernel	GPU Weight Format	GEMM Kernel	MLA Kernel	Siliconflow	Ktrans Point
MMLU (shuffle 1k)							
1	bf16	cpuinfer	bf16	torch	torch	81.6	81.9
2	q8_0	cpuinfer	bf16	torch	torch	81.6	83.1
3	q4km	cpuinfer	bf16	torch	triton	81.6	81.4
4	q4km	cpuinfer	q4km->marlin 8	marlin	triton	81.6	81.1
5	q4km	cpuinfer	q4km->marlin 4	marlin	triton	81.6	81
6	q4km	cpuinfer	fp8	fp8gemm	triton	81.6	81.5
MMLU-pro							
1	q4km	cpuinfer	fp8	fp8gemm	triton	57.7	57.6
2	q4km	cpuinfer	q4km->marlin 4	marlin	triton	57.7	57.5
HumanEval	tbd	tbd	tbd	tbd	tbd	tbd	tbd
GSM8K	tbd	tbd	tbd	tbd	tbd	tbd	tbd

We only sample 1k from 10k MMLU dataset, test once, and do not use few shot, thus the score is lower than the number reported in paper. More will come, updated on Github repo.

Ktransformers v0.2.4： Concurrent Request Support

- Support Continues Batch and Chunked Prefill via an asynchronous architecture learnt from SGLang
- Better total output tokens due to the share parts in GPU





```
(base) rasp@rasp-AvenueCity:~/Projects/ktrans
formers-dev-tmp/ktransformers$ python client.
py --question_id 0
Sure! Here's a longer joke for you:

---

A man walks into a library and asks the libra
rian, "Do you have any books on time travel?"

The librarian looks at him and says, "Yes, bu
t they're all ▯
```

```
(base) rasp@rasp-AvenueCity:~/Projects/ktrans
formers-dev-tmp/ktransformers$ python client.
py --question_id 1
To determine which number is greater between
3.9 and 3.11, we can analyze them step by ste
p. Here's a detailed explanation:

### Step 1: Understand the Numbers
- **3.9**: ▯
```

```
(base) rasp@rasp-AvenueCity:~/Projects/ktrans
formers-dev-tmp/ktransformers$ python client.
py --question_id 2
Certainly! Here's an introduction to the Harr
y Potter series in around 500 words:

---

**Harry Potter: A Magical Journey**

The *Harry Potter* series, written by British
 author J.K. Rowling, is one ▯
```

```
(base) rasp@rasp-AvenueCity:~/Projects/ktran
sformers-dev-tmp/ktransformers$ python clien
t.py --question_id 3
Modern world history, spanning from the late
 15th century to the present, is a narrative
 of profound transformation, marked by explo
ration, revolution, industrialization, and g
lobalization. This era witnessed the rise an
d fall of empires, the █
```

Single concurrency
Decode: 17.6 tokens/s

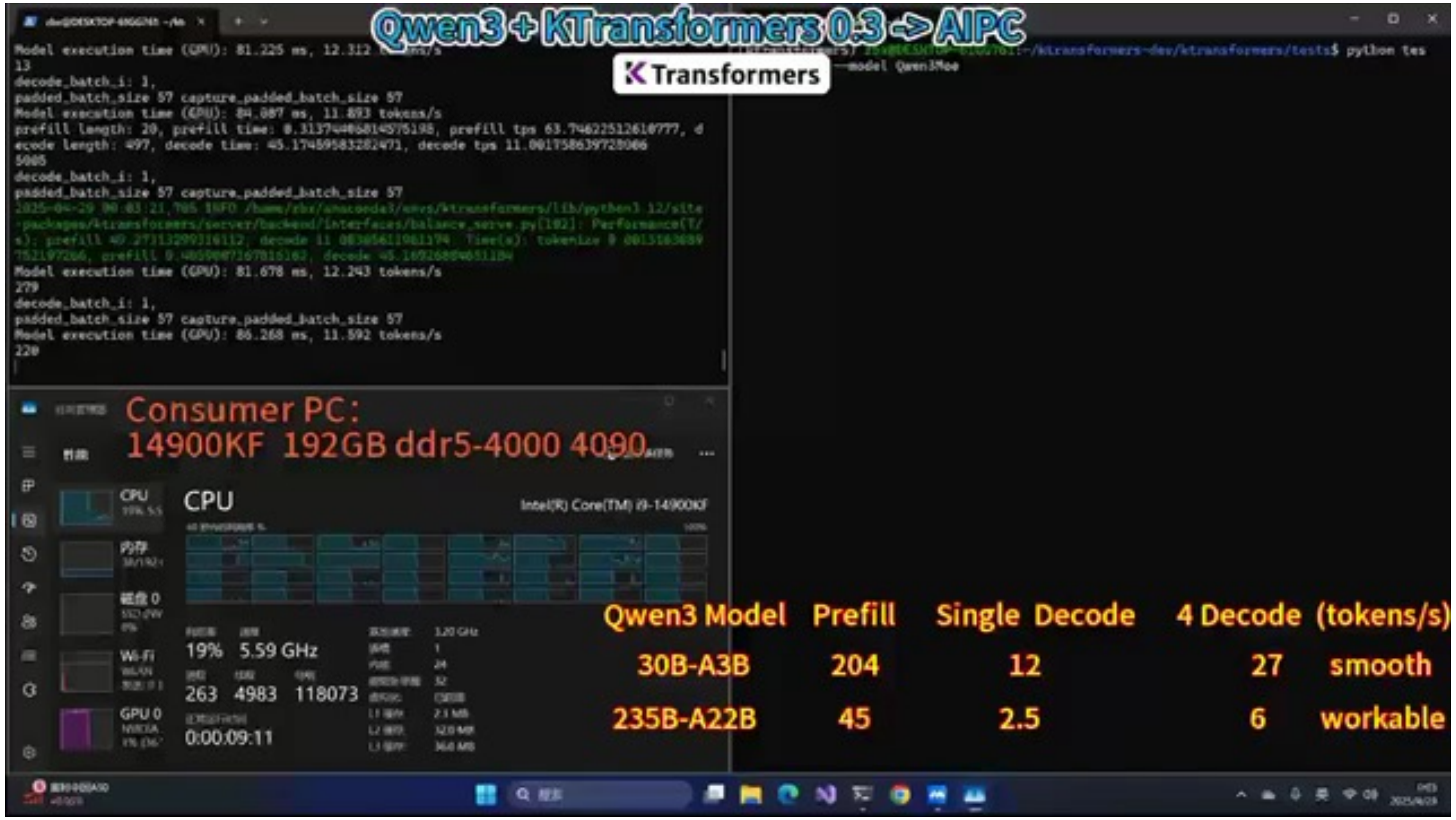
Four-way concurrency
Decode: 10*4 tokens/s

Generate Speedup: 2.27x

DeepSeek-R1/V3-Q4 671B on 4090D 24G + XEON6+ 1T MRDIMM 8800

KTransformers v0.3： Qwen3 Support, the dawn of real AI PC

The post-trained models, such as **Qwen3-30B-A3B**, along with their pre-trained counterparts (e.g., **Qwen3-30B-A3B-Base**), are now available on platforms like **Hugging Face**, **ModelScope**, and **Kaggle**. For deployment, we recommend using frameworks like **SGLang** and **vLLM**. For local usage, tools such as **Ollama**, **LMStudio**, **MLX**, **llama.cpp**, and **KTransformers** are highly recommended. These options ensure that users can easily integrate Qwen3 into their workflows, whether in research, development, or production environments.



Qwen3 + KTransformers 0.3 -> AIPC

KTransformers

```
Model execution time (GPU): 81.225 ms, 12.312 tokens/s
13
decode_batch_i: 1,
padded_batch_size 57 capture_padded_batch_size 57
Model execution time (GPU): 84.087 ms, 11.893 tokens/s
prefill length: 20, prefill time: 0.31374406814575198, prefill tps 63.74622512610777, d
ecode length: 497, decode time: 45.17489583282471, decode tps 11.001758639728006
5005
decode_batch_i: 1,
padded_batch_size 57 capture_padded_batch_size 57
2025-04-29 00:03:21,705 INFO /home/zbx/anaconda3/envs/ktransformers/lib/python3.12/site
-packages/ktransformers/server/backend/interfaces/balance_serve.py[182]: Performance(T/
s): prefill 40.27313299316112, decode 11.083056119611174. Time(s): tokenize 0.0015183089
752197266, prefill 0.49659887167816162, decode 45.169268804651184
Model execution time (GPU): 81.678 ms, 12.243 tokens/s
279
decode_batch_i: 1,
padded_batch_size 57 capture_padded_batch_size 57
Model execution time (GPU): 86.268 ms, 11.592 tokens/s
220
```

```
rs) [illegible]:~/ktransformers-dev/ktransformers/tests$ python tes
 --model Qwen3Moe
```

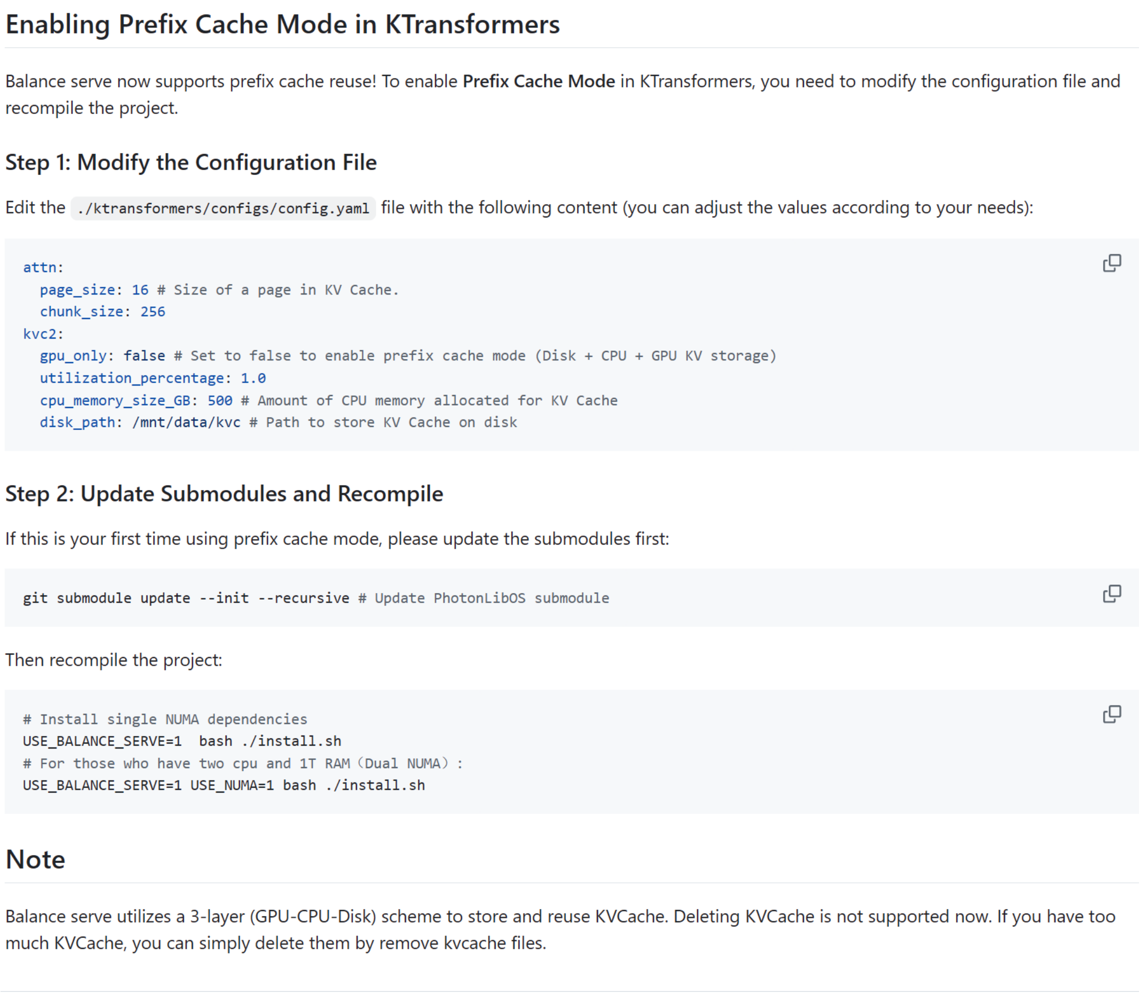
Consumer PC:
14900KF 192GB ddr5-4000 4090

Qwen3 Model	Prefill	Single Decode	4 Decode	(tokens/s)
30B-A3B	204	12	27	smooth
235B-A22B	45	2.5	6	workable

Intel 14900KF + 4090

Still much room for optimization!

KTransformers v0.3.2： L3 Cache and Kimi K2



Enabling Prefix Cache Mode in KTransformers

Balance serve now supports prefix cache reuse! To enable **Prefix Cache Mode** in KTransformers, you need to modify the configuration file and recompile the project.

Step 1: Modify the Configuration File

Edit the `./ktransformers/configs/config.yaml` file with the following content (you can adjust the values according to your needs):

```
attn:
  page_size: 16 # Size of a page in KV Cache.
  chunk_size: 256
kvc2:
  gpu_only: false # Set to false to enable prefix cache mode (Disk + CPU + GPU KV storage)
  utilization_percentage: 1.0
  cpu_memory_size_GB: 500 # Amount of CPU memory allocated for KV Cache
  disk_path: /mnt/data/kvc # Path to store KV Cache on disk
```

Step 2: Update Submodules and Recompile

If this is your first time using prefix cache mode, please update the submodules first:

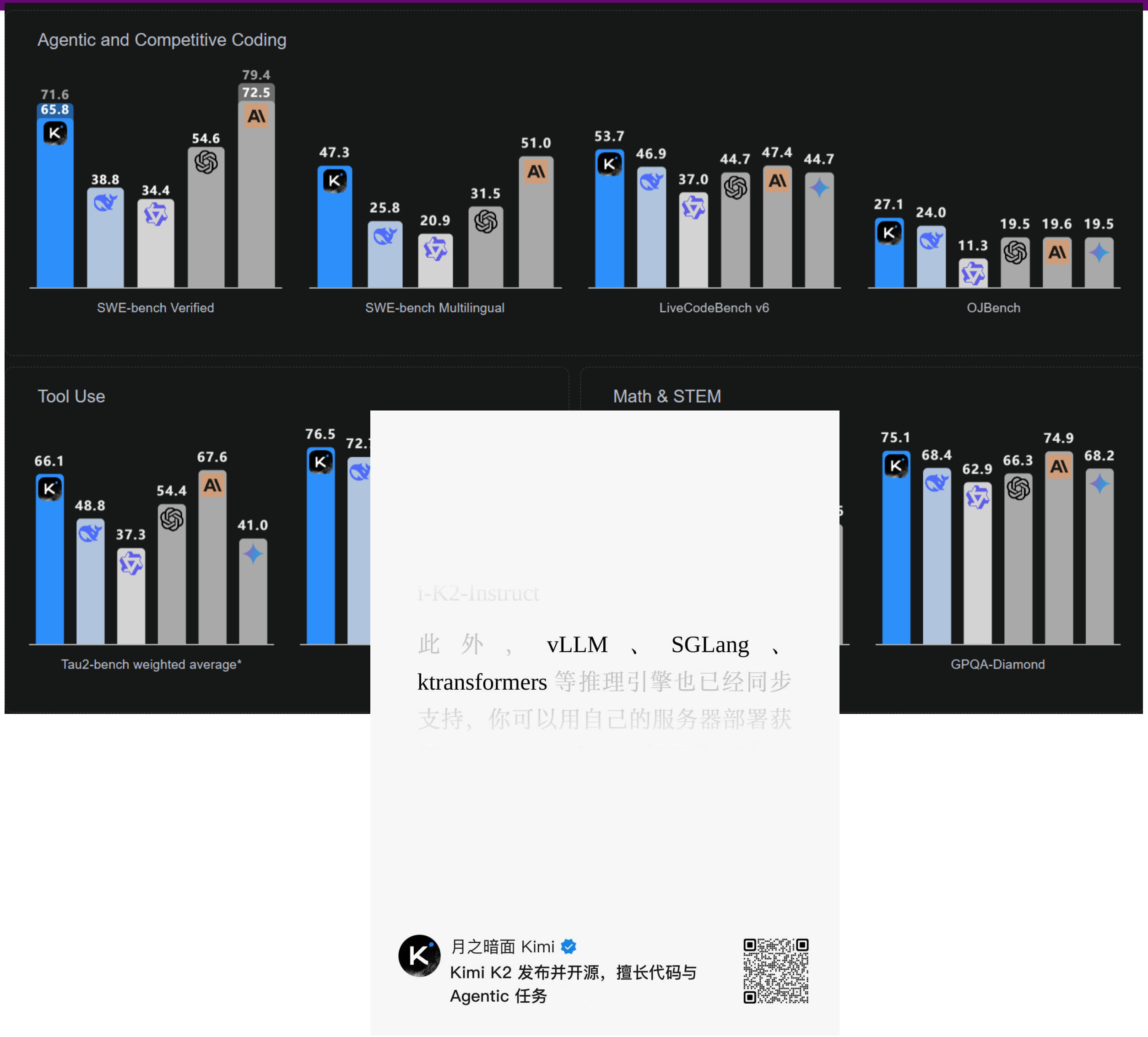
```
git submodule update --init --recursive # Update PhotonLibOS submodule
```

Then recompile the project:

```
# Install single NUMA dependencies
USE_BALANCE_SERVE=1  bash ./install.sh
# For those who have two cpu and 1T RAM（Dual NUMA）:
USE_BALANCE_SERVE=1 USE_NUMA=1 bash ./install.sh
```

Note

Balance serve utilizes a 3-layer (GPU-CPU-Disk) scheme to store and reuse KVCache. Deleting KVCache is not supported now. If you have too much KVCache, you can simply delete them by remove kvcache files.



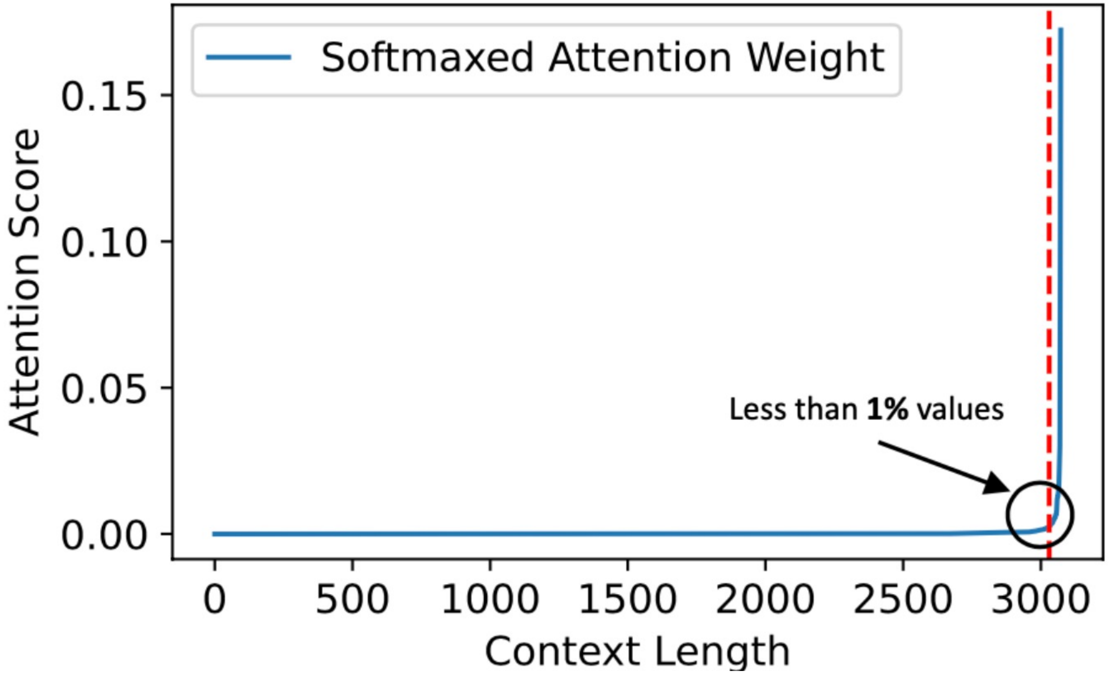
i-K2-Instruct

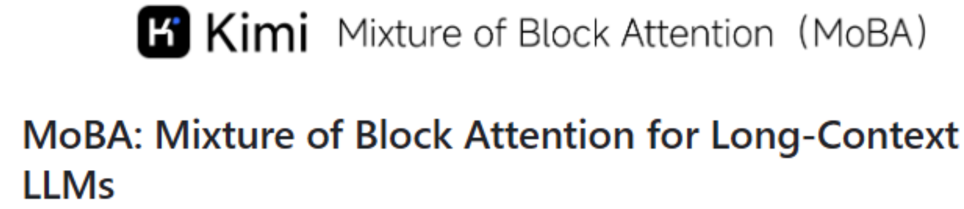
此外，vLLM、SGLang、ktransformers 等推理引擎也已经同步支持，你可以用自己的服务器部署获

月之暗面 Kimi

Kimi K2 发布并开源，擅长代码与 Agentic 任务

Sparsity of FFN/MoE is good, what about the Attention?





Kimi Mixture of Block Attention（MoBA）

MoBA: Mixture of Block Attention for Long-Context LLMs

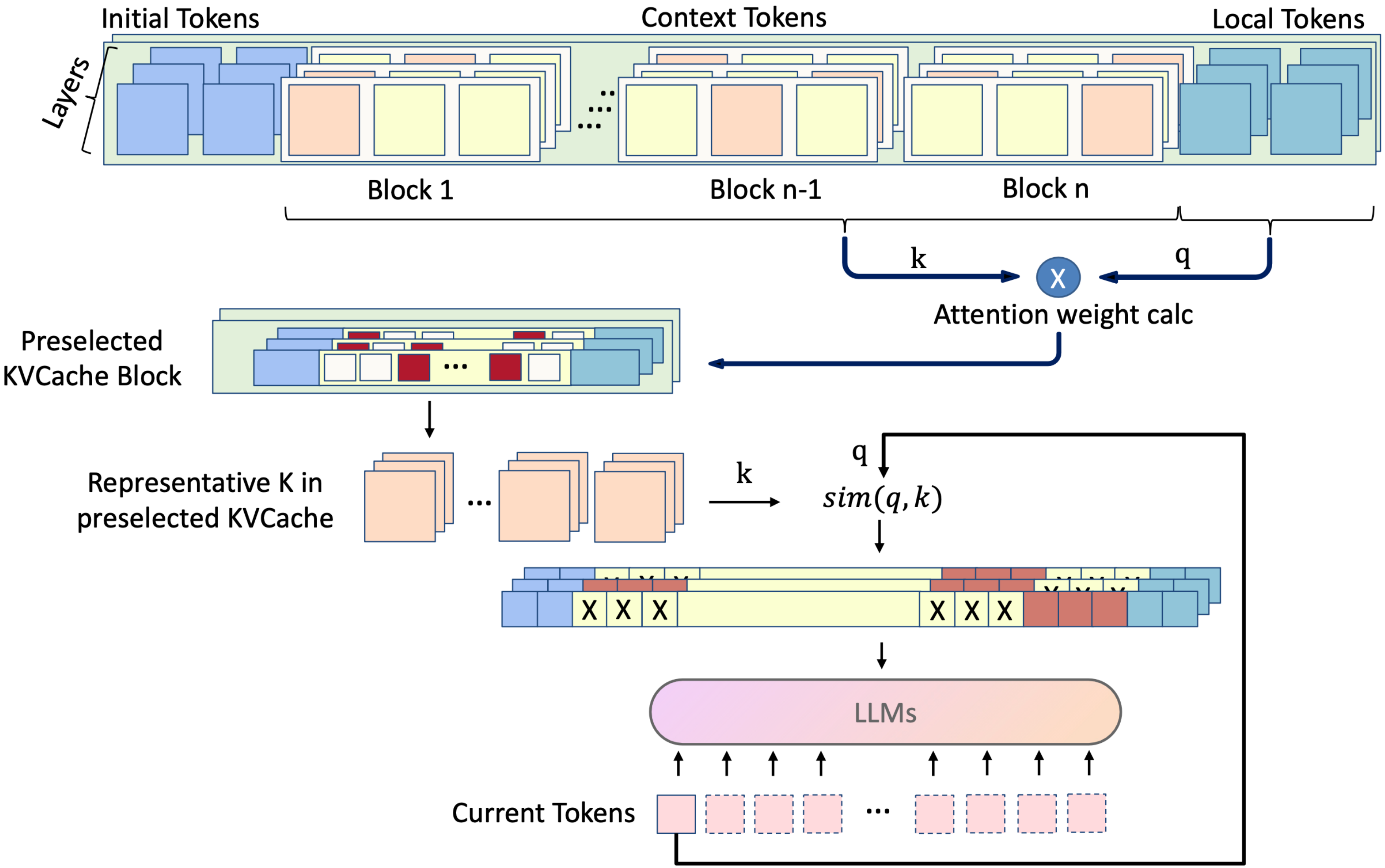
deepseek

Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention

Natural sparsity because of Softmax

Learnable Block-based Sparsity

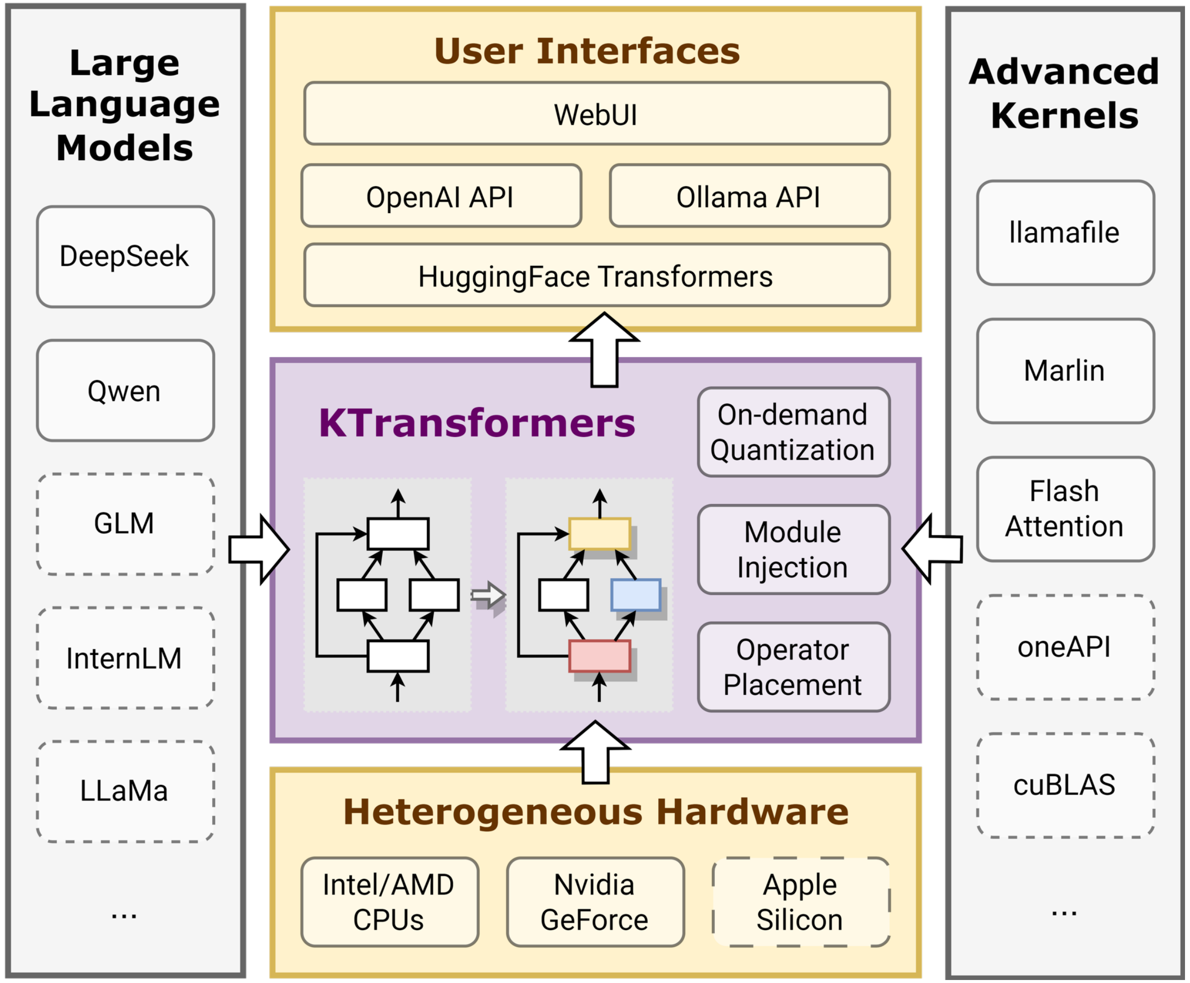
A flexible CPU offload framework for sparse attention

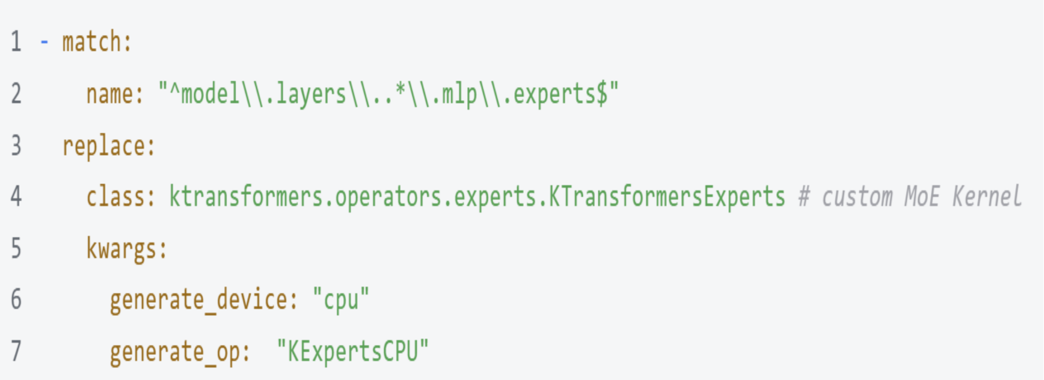


- Initial and local tokens, split into blocks and dynamically selected
- Integrate with Quest, SnapKV, InfLLM, …
- 100% in NIHS <- but this is a weak bmk

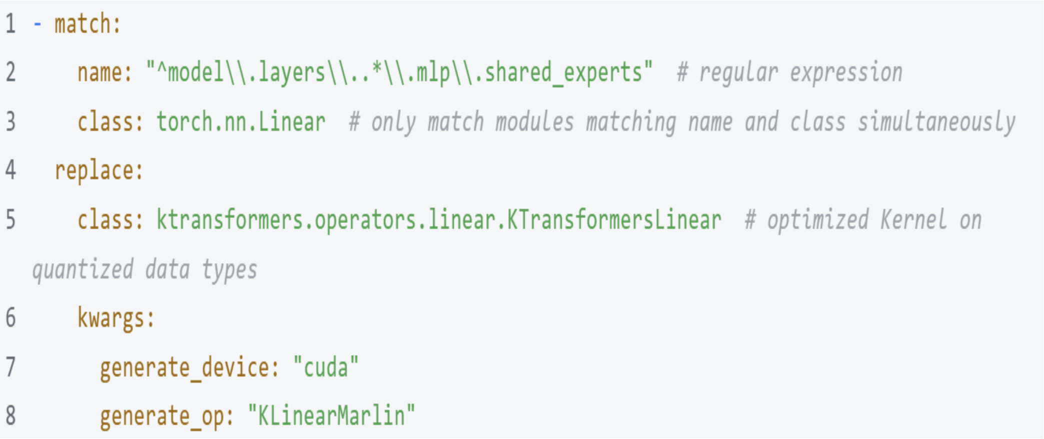
What Ktransformers originally designed for: flexible injection framework

- A combination of single-operator optimizations with each other



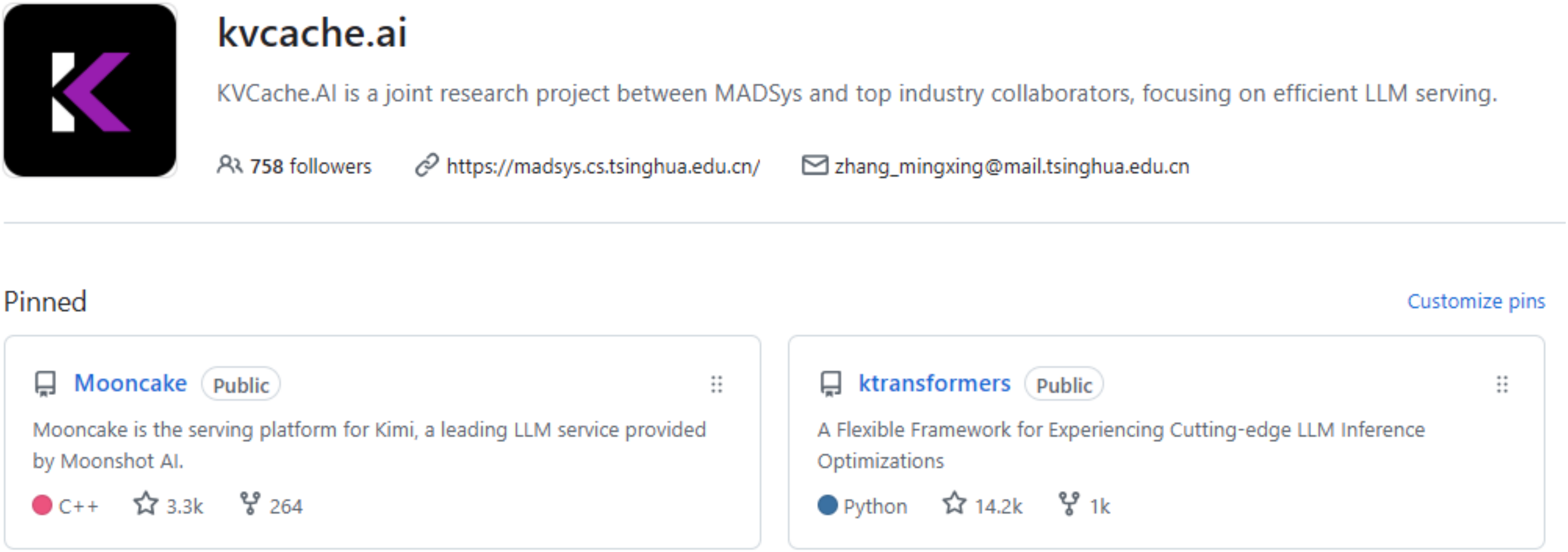


```yaml
- match:
    name: "^model\\.layers\\..*\\.mlp\\.experts$"
  replace:
    class: ktransformers.operators.experts.KTransformersExperts # custom MoE Kernel
    kwargs:
      generate_device: "cpu"
      generate_op:  "KExpertsCPU"
```



```yaml
- match:
    name: "^model\\.layers\\..*\\.mlp\\.shared_experts"  # regular expression
    class: torch.nn.Linear  # only match modules matching name and class simultaneously
  replace:
    class: ktransformers.operators.linear.KTransformersLinear  # optimized Kernel on quantized data types
    kwargs:
      generate_device: "cuda"
      generate_op: "KLinearMarlin"
```

Thanks!



kvcache.ai

KVCache.AI is a joint research project between MADSys and top industry collaborators, focusing on efficient LLM serving.

758 followers https://madsys.cs.tsinghua.edu.cn/ zhang_mingxing@mail.tsinghua.edu.cn

Pinned

Customize pins

Mooncake Public

Mooncake is the serving platform for Kimi, a leading LLM service provided by Moonshot AI.

● C++ ☆ 3.3k 264

ktransformers Public

A Flexible Framework for Experiencing Cutting-edge LLM Inference Optimizations

● Python ☆ 14.2k 1k

https://github.com/kvcache-ai