Memory Operations Tracing in LLM Inference

R244 Presentation

Juntong Deng

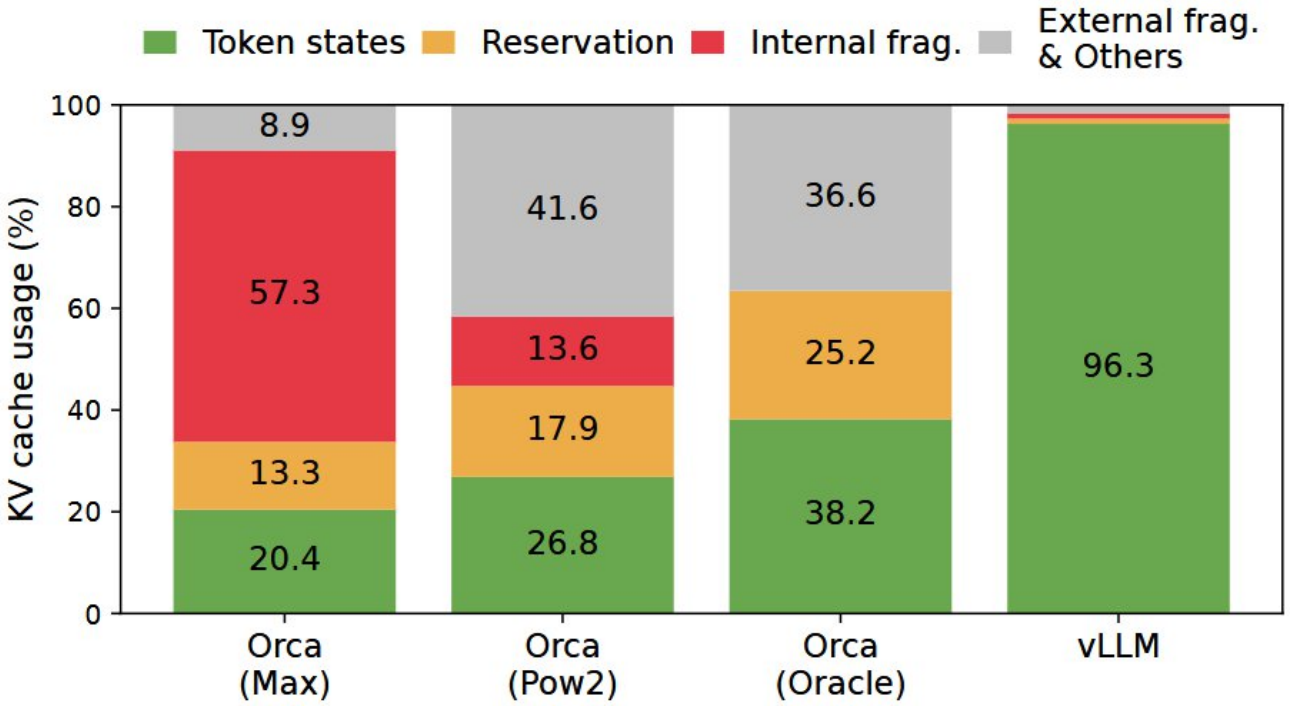
jd2125@cam.ac.uk

Selected Open Source Project

- vLLM [4]
- One of the most widely used LLM inference engines
- Solved memory fragmentation and under-utilization in LLM inference
- PagedAttention [2]: Inspired by virtual memory and paging mechanisms in operating systems
- I focused on operating systems (Linux kernel) throughout my undergraduate studies
- I am very interested in operating systems
- When I see mechanisms from operating systems being applied in machine learning systems,

  I find it very interesting
- I believe that the principles of all systems are interconnected

What does the memory look like in vLLM?



- The author claims that vLLM can achieve a memory utilization rate of 96.3%
- However, this is only a general overview...

What does the memory blocks look like?

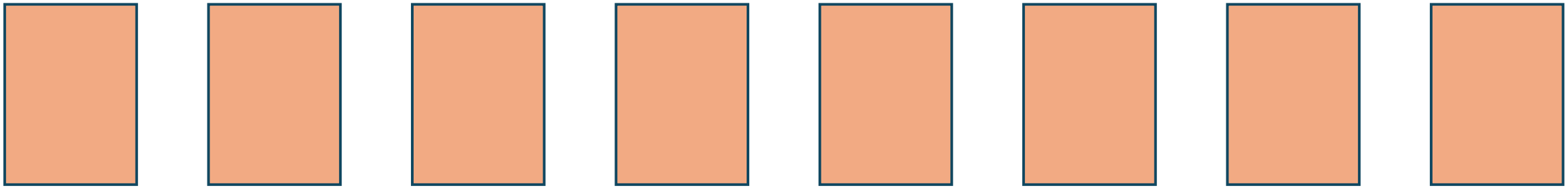
Does it look like this? (utilization rate is uniformly distributed)

Or like this? (utilization rate is randomly distributed)

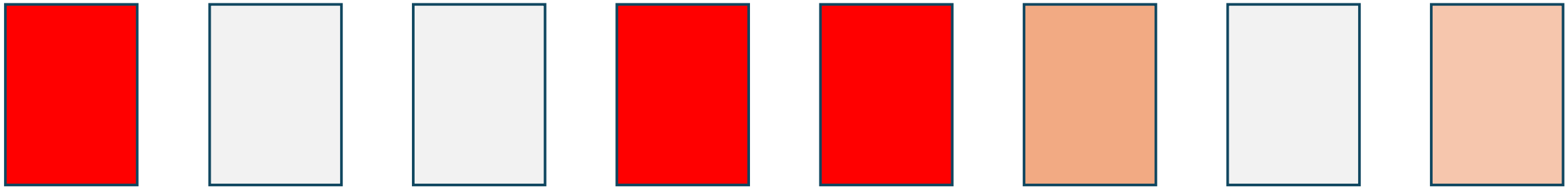
Will it change with different lengths or types of input?

What are memory access patterns?

Does it look like this? (the access frequency of each memory block is close)



Or like this? (there are hot or cold memory blocks)



Will it change with different lengths or types of input?

Hard to know...

vLLM does not support tracing the state and operations of memory blocks...

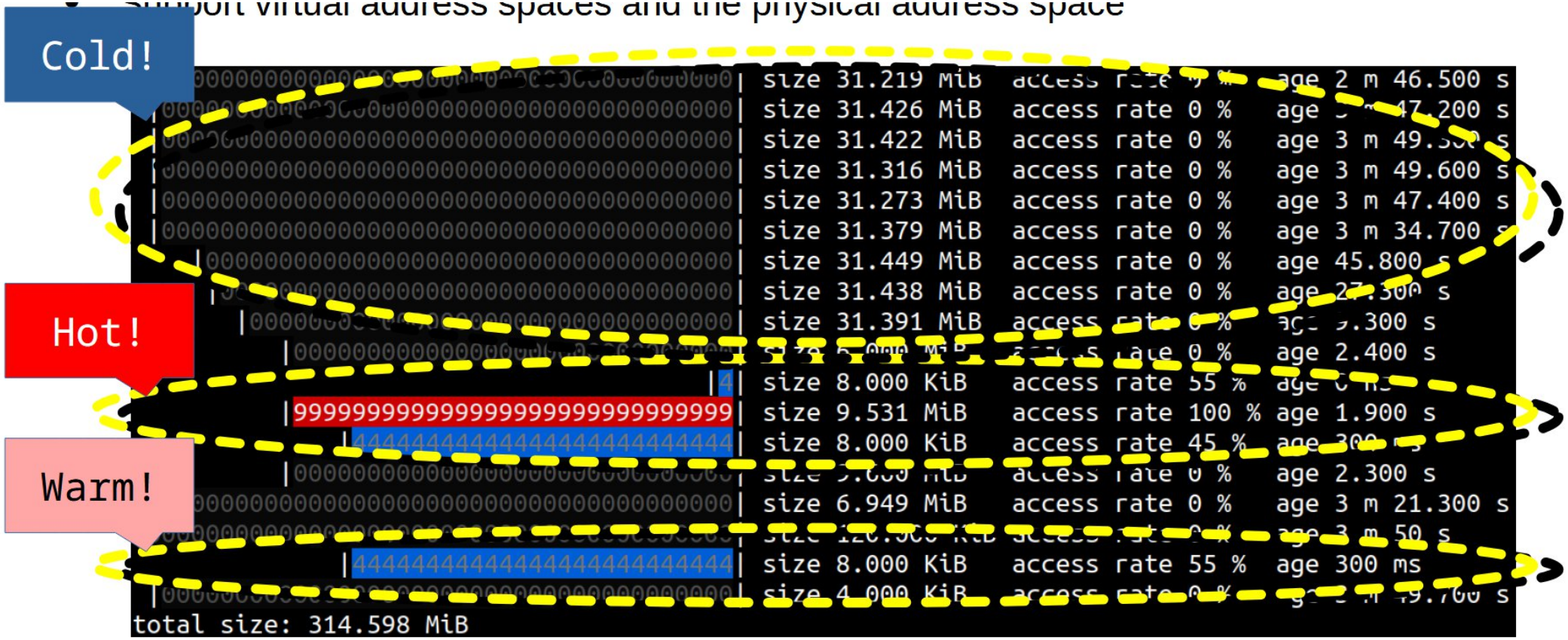
Observability

- Observability is very important in complex systems
- If we don't know what is happening...
- How can we optimize the system?

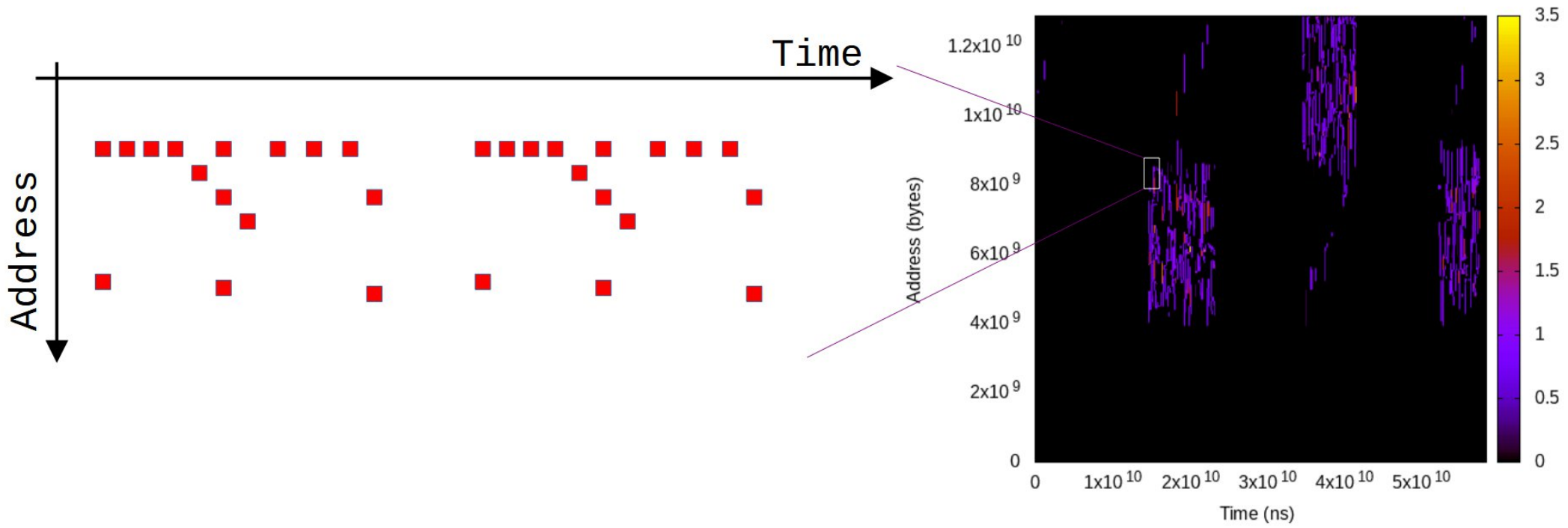
My mini project

- Add memory block operation tracing to vLLM
- Let what happened to memory blocks in LLM inference visible
- Help analyze memory utilization distribution and memory access patterns in LLM inference
- Help further memory optimization in LLM inference
- Inspired by memory access monitoring in operating systems
- DAMON: Data Access MONitor (Linux kernel) [1] [3]
- Not only for optimization, it would also be interesting to visualize the memory state and operations in LLM inference (can help understand how vLLM works)

DAMON Screenshot



DAMON Screenshot



Progress?

- To be honest...
- I have been very busy recently and really do not have time...
- Sorry, but I will start soon!
- Plan?
- I plan to complete it within this month (hopefully...)

Thanks

Thank you for listening

References

[1] DAMON. 2019. DAMON: Data Access Monitor. https://damonitor.github.io/. Accessed on Dec 2, 2025.

[2] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In Proceedings of the 29th Symposium on Operating Systems Principles (Koblenz, Germany) (SOSP '23). Association for Computing Machinery, New York, NY, USA, 611–626. doi:10.1145/3600006.3613165

[3] SeongJae Park, Yunjae Lee, and Heon Y. Yeom. 2019. Profiling Dynamic Data Access Patterns with Controlled Overhead and Quality. In Proceedings of the 20th International Middleware Conference Industrial Track (Davis, CA, USA) (Middleware '19). Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3366626.3368125

[4] vLLM. 2025. Welcome to vLLM. https://docs.vllm.ai/en/latest/. Accessed on Dec 2, 2025.