Event Info

https://github.com/LMCache/LMCache

Zoom link

Time: 9am–9:30am PT every other week, starting from Jul 29, 2025

Recordings

https://www.youtube.com/@LMCacheTeam

Oct 6th

- Attendees:
    - Jiayi Yao (LMCache Lab)
    - Shaoting Feng (LMCache Lab)
    - Samuel Shen (LMCache Lab)
    - Dongjoo Seo (Samsung)
    - Luis Chamberlain (Samsung)
    - Mo McElaney (IBM)
    - Hasan Arif (Virginia Tech)
- Discussion items
    - (Kobe) Documentation Efforts
    - (Luis) KV cache growth projections and kv cache simulator (kvcache.io)
        - Include newer models
        - Can include in documentation, in kv cache calculator, or a separate project under lmcache
        - https://github.com/LMCache/kvcache-view
            - Luis will manage this as admin
    - (Shaoting) More e2e tests
    - (Jiayi) Refactor centralized sharing code (with transfer channel) such that RDMA can be leveraged
    - (Jiayi) Will start working on fault tolerance soon
    - (Mo) LMCache office hours - topics - Potential first call Oct 30 at 11am PT

Sep 23rd

- Attendees:
    - Dongjoo Seo (Samsung)
    - Luis Chamberlain (Samsung)
    - Yihua Cheng (LMCache Lab)
    - Jiayi Yao (LMCache Lab)
    - Arivappa
    - Walter Beller-Morales (Cohere)
- Discussion items
    - (Martin) Relevant unit tests can now run on non-CUDA environment (https://github.com/LMCache/LMCache/pull/1575)
        - GitHub workflow for running unit tests
        - Next stage: Add check that if the tests fail on GitHub don't run tests on GCP (Buildkite)
    - (Martin) Extend Backend refactored to use new interface `ConfigurableStorageBackendInterface` to provide definitive constructor signature for implementers (https://github.com/LMCache/LMCache/pull/1636):

        - Near to merge, need another +1
        - Next stage: Suggestion to rename external backends to configurable backends. What do people think?
    - (Martin) GDS backend tests were not running:
        - PR to fix this: https://github.com/LMCache/LMCache/pull/1648
        - Additional PR to enable it to run on CI on GCP: https://github.com/LMCache/LMCache/pull/1659
    - (Dongjoo) Improving H2D memory copying by hugepage
    - (Jiayi) RDMA-based P2P KV Cache sharing based on nixl: https://github.com/LMCache/LMCache/pull/1610
    - (Yihua/Jiayi) Refactor the PD disagg part with configuration changes
        - https://github.com/LMCache/LMCache/pull/1579

Sep 9th

- Attendees:
    - Yihua cheng (LMCache Lab)
    - Shaoting Feng (LMCache Lab)
    - Dongjoo Seo (Samsung)
    - Caesar Chen (AWS Sagemaker)
    - Chandra Lohit Reddy Tekulapally (AWS Sagemaker)
    - Yanlin DU (Penn State Uni.)
    - Arivappa
- Discussion items
    - (Jiayi) Async KV cache loading: https://github.com/LMCache/LMCache/pull/1513
    - (Yihua) Pytest benchmark for engine lookup store and retrieve operations: https://github.com/LMCache/LMCache/pull/1484
    - (Shaoting) Disaggregated prefill bugfix:
        - https://github.com/LMCache/LMCache/pull/1494
        - https://github.com/LMCache/LMCache/pull/1446
        - https://github.com/LMCache/LMCache/pull/1472
    - (Samuel) Mock connector: https://github.com/LMCache/LMCache/pull/1500
    - (Chandra / Caesar) AWS SageMaker AI toolkit connector https://github.com/LMCache/LMCache/pull/1111
    - (Dongjoo) Purpose for Hugepage / THP? What architecture is currently mainly targeting?([Enhancement] Add hugepage memory support by DongDongJu · Pull Request #1442 · LMCache/LMCache)

Aug 26nd

- Recording:
- Attendees:
    - Dongjoo Seo (Samsung)
    - Yuhan Liu (LMCache Lab)
    - Yihua Cheng (LMCache Lab)
    - Jiayi Yao (LMCache Lab)
    - Baoloongmao(Tencent.inc)
    - Arivappa
- Discussion items
    - (Yihua) New interface for async `lookup` operation at vLLM side: https://github.com/vllm-project/vllm/pull/23620
        - 2 potential ways to use this interface
            - Reduce the lookup overhead in the scheduler process
            - Enable KV cache prefetching
    - (Dongjoo) Support for CXL memory within app-level memory management
        - Needed feature
            - CPU-level NUMA control (https://github.com/LMCache/LMCache/pull/1417)
            - Memory migration policy for tiered memory. Maybe extend the disk swapping policy or combine with that.
            - Link to slack channel: https://join.slack.com/t/lmcacheworkspace/shared_invite/zt-3c5mzct0k-zeGla1ja9UPq~tR0H9MoIg
    - (Baoloongmao) https://github.com/LMCache/LMCache/pull/1439 Support dynamical load external backend
        - Documentation update for the feature?
    - (Jiayi) Amazon S3 remote connector
    - (Jiayi) Cacheblend ready for testing
        - Fixed several bugs
        - Add sparse attention to generate accurate attention mask, leading to better accuracy
        - Plan for supporting DeepSeek?
    - (Arivappa) Support for ARM64 https://github.com/LMCache/LMCache/issues/800
        - We have mainly two dependencies, nixl and infinistore packages, which don't support ARM right now.
        - I have tried to build arm64 for infinistore, it worked - need to followup.
        - (community) Maybe we can try the new feature dynamic linking to support ARM (avoiding the dependency of nixl and infinistore .. ?)
        - As I don’t have much experience in this area, I would love to receive any feedback or guidance, thank you.

Aug 12nd

- Recording: https://youtu.be/r4bzaXr1JWc
- Attendees:
    - Yuhan Liu (LMCache Lab)
    - Shaoting Feng (LMCache Lab)
    - Samuel Shen (LMCache Lab)
    - David Edelsohn (NVIDIA)
- Discussion items
    - [CI/CD] Have a comprehensive unit test that is only triggered when the PR is tagged as “ready”: https://github.com/LMCache/LMCache/pull/1249
        - Top features that need unit test:
            - CacheBlend
            - PD
            - Cache Controller
            - Layerwise
    - [Community] Rule for new committers
        - We have rules for github actions, but not buildkite
        - Need documentation for the rules
    - [New architecture] Separated process LMCache: https://github.com/vllm-project/vllm/issues/22605
    - [New architecture] LMCache C++/Rust backend
    -

July 29th

- Recording: https://youtu.be/ThWcWC0pKIU
- Attendees:
    - Martin Hickey (IBM)
    - Yuhan Liu (LMCache Lab)
    - Baolong Mao (Tencent)
    - Jiayi Yao (LMCache Lab)
- Discussion items:
    - [Martin] Add an OWNERS file to the repo?
    - [Martin] Some review best practices:
        - PR description should be useful to the reader
        - Wait to merge until CI passes
    - [Martin] Timeline to change default branch from ‘dev’ to ‘main’
    - [Yihua] Documentation about new LMCache configs (and `extra_config`)

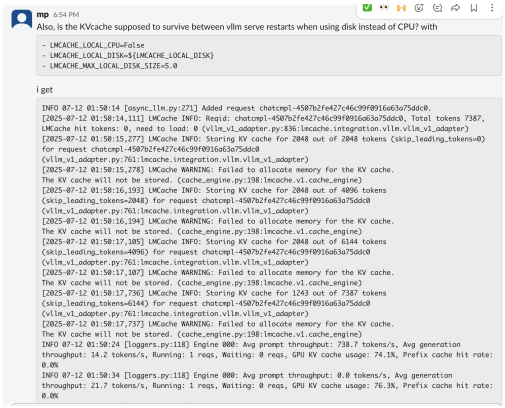
    - [Martin] Policy for supporting different vLMM versions?
        - Ref: https://github.com/LMCache/LMCache/issues/1073#issuecomment-3130559351
        - We should document the versions that a release support
    - [Martin] Looking for review of following PRs:
        - https://github.com/LMCache/LMCache/pull/1145
        - https://github.com/LMCache/LMCache/pull/1087
    - [Baoloong] Refactor to adapt multiply Inference engine
    - [Jiayi] Optimize disk and prefetch code path
    - [Jiayi] [WIP] Support external cross-node move/compress/decompress operations (issued by operators)
    - [Baoloong] supplies a way to test lmcache without vllm.

July 22nd

No meeting this week. Meetings now changed to bi-weekly starting next Tuesday (29th) at 09:00 am PT.

July 15th

- Recording:
- Attendees:
    - Yuhan Liu (LMCache Lab)
    - Shaoting Feng (LMCache Lab)
    - Yihua Cheng (LMCache Lab)
    - Nick Barcet (nijalabs)
    - Kuntai Du (vLLM & LMCache)
    - Samuel Shen (LMCache Lab)
    - Kai-Hsun Chen (Anyscale)
- Discussion items:
    - [Shaoting] Tensor parallel support for XPYD (disaggregated prefill)
        - Ref: https://github.com/LMCache/LMCache/pull/1039, https://github.com/LMCache/LMCache/pull/1044
        - Future roadmap: heterogeneous tensor parallelism support
    - [Samuel] Memory Leak https://github.com/LMCache/LMCache/pull/1046
        - Related issues:



    - https://github.com/LMCache/LMCache/issues/1034
    - https://github.com/LMCache/LMCache/issues/955
  - [Yihua] Roadmap for PD disagg and CPU offloading
    - Current status: PD disagg and CPU offloading are using different code paths, and cannot be activated at the same time
    - Plan: aggregate the code paths and make it being able to activate at the same time
    - Timeline: a few weeks

July 8th

- Recording: LMCache Community Meeting 07/08
- Attendees:
  - Yuhan Liu (LMCache Lab)
  - Jiayi Yao (LMCache Lab)
  - Riggins (Tencent)
  - Mo McElaney (IBM)
  - Martin Hickey (IBM)
  - RuiQi (NTU Singapore)
  -
- Discussion items:
  - [Martin] vLLM integration testing:
    - Ref: https://github.com/LMCache/LMCache/issues/920
      - Eventually we want multiple requests to test it
    - https://github.com/LMCache/LMCache/pull/930
      - Ready to be merged
  - [Mo] LMCache Community Outreach plan (still in progress)
  - [Riggins] Any update on error handling? https://github.com/vllm-project/vllm/pull/19330
  - [Martin]: Mix of html and markdown. Do we want to do this?

    - [Martin] Going to triage old PRs and issues. Do we need a “stale-bot” in addition?
        - Github actions: if no activities in an issue/PR after 90 days, it will automatically close the issue/PR
    - [RuiQi] xPyD Bug (https://github.com/LMCache/LMCache/issues/981)

July 1st

- Recording: https://youtu.be/7Rnl1bzd3a8
- Attendees:
    - Yuhan Liu (LMCache Lab)
    - Riggins (Tencent)
    - Jiayi Yao (LMCache Lab)
    - Eric Shan (UChicago / LMCache Lab
    - Baoloongmao (Tencent)
- Discussion items:
    - [Jiayi] XPYD ready to be tested: https://github.com/LMCache/LMCache/pull/895
    - [Jiayi] Several bugfixes
        - Working on the optimization of overhead CPU
    - [Riggins] skip_last_n_tokens bugfix https://github.com/LMCache/LMCache/pull/942
    - [Baoloongmao] Accuracy declined while enable discard_partial_chunks https://github.com/LMCache/LMCache/issues/945

June 24th

- Recording:
- Attendees:
    - Yuhan Liu (LMCache Lab)
    - Riggins (Tencent)
    - Martin Hickey (IBM)
    - Mo McElaney (IBM)
    - Baoloongmao(Tencent)
    - Jiayi Yao (LMCache Lab)
    - Eric Shan (UChicago / LMCache Lab)
    - Dan Aloni (VAST Data)
- Discussion items:
    - [Jiayi] XPYD support: will be ready soon https://github.com/LMCache/LMCache/pull/895

    - [Shaoting] Multimodal support (image models) https://github.com/LMCache/LMCache/pull/882
    - [Yuwei] SGLang support https://github.com/LMCache/LMCache/pull/869
    - [Martin] Container image will now be built automatically when a release is cut (https://github.com/LMCache/LMCache/pull/784) Also, nightly builds of image (https://github.com/LMCache/LMCache/pull/756).
    - [Riggins] Fallback strategy when disk load fails
        - Got the number of matched tokens and then start loading. But what if there is power outage / bugs, there's no try catch mechanism
            - Ongoing effort on this -> PR regarding error handling, talking to AWS on error handling and will be fixed once PR is merged.
    - [Riggins] Skip last n tokens bug fix. https://github.com/LMCache/LMCache/pull/897
    - [Martin] Static checker for GH workflows. Ref: https://github.com/LMCache/LMCache/pull/808
    - [Martin] Improve test unit coverage metrics. @Jiayi Are we going in the direction you envisage? Ref: https://github.com/LMCache/LMCache/pull/823
    - [Mo] Who would I talk to about what the LMCache community needs help with?
    - [Baoloongmao] Just two PRs need to be reviewed
        - https://github.com/LMCache/LMCache/pull/858
        - https://github.com/LMCache/LMCache/pull/764

June 17th

- Recording:
- Attendees:
    - Yuhan Liu (LMCache Lab)
    - Yihua Cheng (LMCache Lab)
    - Baoloongmao (Tencent)
    - Eric Shan (UChicago / LMCache Lab)
    - Yuwei An(LMCache Lab)
    - Kuntai Du (LMCache Lab)
    - Kobe Chen (LMCache Lab)
- Discussion items:
    - [Jiayi] Performance optimizations:

- Layer-wise pipelining to reduce the overhead for KV cache store (from 5% to 2-3%)

    - [In progress] MLA support with PD
  - PD proxy optimization to reduce the overhead in TTFT
    - 1P1D setup, 8K input, 200 output, Llama 8B model → reduce TTFT from 425ms to 310ms
- [Jiayi/Yihua] XpYd in progres, PR in this week and early next week
- [Yihua] Feature compatibility testing and improvement
  - Features: MLA, PD, CPU offloading, Layerwise pipeline
- [Baoloongmao]
  - https://github.com/LMCache/LMCache/pull/858
    - (858) need reviewer, develop new remote connector developer without modifying existing code—reduce conflict and do not need dependency of old developer. Will finish this PR (previous updates have not merged, and can bypass the review process)
  - https://github.com/LMCache/LMCache/pull/843
  - https://github.com/LMCache/LMCache/pull/764
  - Why we upgrade to the torch version from 2.6 to 2.7? It there a way to build a lmcache with torch 2.6? Since we are using a last month built vllm which is dependent torch 2.6
    - [Yihua] Change the torch version in `pyproject.toml` or use `pip install `--no-build-isolation`

- [Kuntai]
  - [News] Incoming PR: Hybrid memory allocator for sliding window in vLLM (PR) that changes vLLM KV cache layout.
  - Investigating using LMCache for prefill-only optimizations in vLLM (design doc)
  - Investigating the potential of KV cache compaction
- [Yihua] Discussion: separate LMCache to another process for offloading and KV transfer
  - Pros:
    - Less impact to vLLM worker processes

                - Directly share KV across different vLLM on the same node without “lmcache server”
                - Reduce the communication overheads across TP workers
                - Reduce the time to bootstrap
            - [Kuntai]: how do we interact with CUDA graph in this design?
            - [Kuntai]: should the scheduler connector be located in the same process as vLLM?
        - [Yuwei] Integration with SGLang is ongoing
        - [Kuntai] vLLM is supporting P2P NCCL-style communication
            - Benefit: no need to configure the connection at launch time -- the connection can be built at runtime.

June 10th

- Recording:
- Attendee:
    - Zhuohan Gu (UChicago / LMCache Lab)
    - Yuhan Liu (UChicago / LMCache Lab)
    - Martin Hickey (IBM)
    - Jonas Rosland (VAST Data)
    - Eric Shan (UChicago / LMCache Lab)
    - Jiayi Yao (UChicago / LMCache Lab)
    - Wei Cai (UChicago)
- Discussion items:
    - [Martin] Nightly container image build of latest LMCache integrated with latest vLLM (https://github.com/LMCache/LMCache/pull/756)
    - [Martin] CUDA wheels build fixed (https://github.com/LMCache/LMCache/pull/775)
    - [Jiayi] Merged CacheBlend PR (https://github.com/LMCache/LMCache/pull/762)
    - [Jiayi] Optimized layer-wise pipelining (https://github.com/LMCache/LMCache/pull/794)
    - [Jiayi] Optimized pd proxy (will be a PR soon)

June 3rd

- Recording:
- Attendee:
    - Yuhan Liu (UChicago / LMCache Team)
    - Jiayi Yao (UChicago/LMCache Lab)
    - Baoloongmao (Tencent)
    - Kuntai Du (UChicago / LMCache Team)
    - Anukul
    - Tyler Michael Smith (Red Hat)
- Discussion items:
    - [Baolong] https://github.com/LMCache/LMCache/pull/742
        - Simple PR that support set extra config in LMCacheEngineConfig
        - Make it easier for developers to specify new configs
    - [Baolong] https://github.com/LMCache/LMCache/pull/764
        - probe thread for remote connector
        - Enabling falling back to recompute performance
    - [Jiayi] https://github.com/LMCache/LMCache/pull/762
        - CacheBlend almost ready for latest LMCache and vLLM v1
    - [Jiayi] Minor bug fixes
    - [Kuntai] Current cachegen degrades accuracy on some models (https://github.com/LMCache/LMCache/issues/754)
    -

May 27th

- Recording:
- Attendee:
    - Yuhan Liu (UChicago / LMCache Team)
    - Zhuohan Gu (UChicago / LMCache Lab)
    - Jiayi Yao (UChicago / LMCache Lab)
    - Rain Jiang (Bytedance)
    - Kosseila (CloudThrill)
    - Hussain (University of Chicago / LMCache)
    - Martin Hickey (IBM)
    - Siddhant Ray (UChicago/ LMCache Lab)
- Discussion items:
    - [Jiayi Yao]
        - Refactor
        - Disagg prefill
        - LMCache connector
        - Cacheblend
    - DCO [Martin Hickey]
    - Pre-commit [Martin Hickey]
        - Code quality check. PR #691. Instructions on Slack #general channel
        - Martin: DCO, standard contribution agreement
        - Kosseila: Issue created #462 around adding static security scan tool to pre-commit hooks (checkov)
    - [Samuel]
        - Benchmarking platform
            - Layer-wise, KV aware, prefix aware
    -

May 20th

- Recording:
- Attendee:
    - Yuhan Liu (LMCache Team / UChicago)
    - Jiayi Yao (LMCache Team/ UChicago)
    - Zhuohan Gu (LMCache Team / UChicago)

    - Kosseila (CloudThrill)
    - Wenlong Wang (Incoming Meta)
    - Samuel Shen (LMCache Team / UChicago)
- Discussion items:
    - Issues are labelled now! e.g. Good First Issue
    - [Wenlong Wang]: latest Docker tag is previous version of vllm (v0)
        - Currently nightly lmcache/vllm-openai image build is in progress
            - https://github.com/LMCache/LMCache/issues/594
        - Issue is described at: https://github.com/LMCache/LMCache/issues/686
    - [Jiayi]
        - Fixing bugs in async layer-wise loading
        - Benchmarking P/D disaggregation
            - XpYd? Currently only Xp1d supported, but XpYd will be supported soon
            - How to decide which baseline to test against?
                - e.g. 3P1D, we will test against 4 regular vllm engines
            - Native KV transfer in vLLM has weak performance
        - Trying to support asymmetric TP size for PD
        - Discussion on KV cache controller to improve interface
        - Collaboration with Mooncake: active PRs to improve Mooncake backend performance
    - [Samuel]
        - Benchmarking repo: https://github.com/LMCache/LMBench
    - [Kosseila]
        - Are we going to put the production stack helm chart onto Artifact Hub (DevOps de facto looks here)
        - Feel free to raise an issue, thanks!

May 13th

- Recording:
- Attendee:
    - Yuhan Liu (LMCache Team/UChicago)
    - Zhuohan Gu (UChicago / LMCache Lab)
    - Jiayi Yao (UChicago/LMCache team)
    - Siddhant Ray (UChicago/LMCache team)
    - Pratik Doshi (UC Santa Cruz/Independent Collaboration)
    - Kosseila (CloudThrill)
    - Baoloongmao(Tencent)
- Discussion items

    - [Jiayi]
        - Offloading / sharing
            - Layer-wise async cache store and retrieve for CPU offloading & PD
        - PD disaggregation
            - XPYD
    - [Baolong]
        - PR #506
        - PR #428 (only v0 for now, test v1 in the future)
    - [Yihua]
        - XPYD (PR is ready)
        - vLLM has a new connector interface for better async execution, integrates with LMCache
    - [Guopeng]: Prefetch implementation in vLLM v1

May 6th

- Recording:
- Attendee:
    - Yuhan Liu (LMCache Team / UChicago)
    - Kuntai Du (vLLM / LMCache / UChicago)
    - Shaoting Feng (UChicago / LMCache Lab)
    - Jiayi Yao (UChicago / LMCache Lab)
    - Huibing Dong (incoming @ Google)
    - Wenlong Wang (Incoming Meta)
    - Kosseila (CloudThrill)
    - Rain Jiang (Bytedance)
    - Baoloongmao (Tencent)
- Discussion items
    - [Jiayi] Improving perf of CPU offloading
        - Layer-wise pipelining
    - [Jiayi] CacheBlend on vLLM v1
    - [Kuntai] PD disagg back pressure, serde improvement
        - [Kuntai] potential deadlock condition
    - [Yihua] Xpyd support in LMCache
        - Current functionalities
            - NIXL-based
            - Can run with tp=1
            - Can build connection dynamically
        - Need bug fix to ship the PR
    - [Rain] MLA
    - [Rain] Saving the hidden state (#368)
        - [Kuntai] this is for v0
        - [Kuntai] for v1 & skip sampling

    - [Yihua] correctness CI, a nightly benchmark (just using the Llama models for now)
    - [Yihua] Dynamo.
    - [Baolong] MLA
        - Some detailed issues — some of the fields in MLA metadata, want to talk to Jiayi w.r.t. the hidden states and MLA metadata-related stuff.

April 29th

- Recording: https://youtu.be/OxhKSKgliNQ
- Attendee
    - Yuhan Liu (vLLM Production Stack Team / UChicago)
    - Zhuohan Gu (LMCache Lab / UChicago)
    - Jiayi Yao (LMCache Lab / UChicago)
    - Rain Jiang(Bytedance)
    - Baoloong Mao(Tencent)
    - Derek (Huawei)
    - Huibing Dong(Google)
    - Jing Zhou (Jaguar Land Rover)
    - David(Meta)
    - Lei Zhang (Bytedance)
    - Lifan(Meta)
    - Meggie (Bytedance)
    - Siddhant Ray (LMCache Lab/ UChicago)
    - Tao Zhang (USTC)
    - Tong (Google)
    - Wenlong Wang (Meta)
- Summary
- Discussion items
    - LMCache Controller: (Jiayi)
        - KV Cache aware routing
        - Load Balancing
        - /lookup
        - /pin
        - /move
    - LMCache for Disagg Prefill
        - 1P1D -> XPYD

- Mooncake Connector: PR # 428（Baoloong）
  - [support KV transfer of MLA backend #428]
  - https://github.com/LMCache/LMCache/pull/428

3. do some small modification to make partition match possible.

Remove this following code within `vllm_adapter.py`

```
    is_all_prefill = all(
        [status == RetrieveStatus.PREFILL for status in retrieve_status])

    if is_all_prefill and num_request_not_found == 0:
        return model_input, False, None
```

- Have we tested chunked prefill with nixl integration? (Rain)
  - vllm v1 chunked prefill is on by default (Jiayi)
- []