Taejun Yoon

Dept. of Computer Science and Engineering, POSTECH, South Korea
Homepage: ytj0604.github.io
Advisor: Wook-Shin Han

Data Systems Lab @ POSTECH
Email: tjyoon@dblab.postech.ac.kr
GitHub: github.com/ytj0604

Research Interests

Caching and Memory Management for Cloud Analytics and LLM Serving

- **Cloud Data Systems:** Designing and evaluating caching strategies for cloud OLAP systems that query data directly from object storage
- **LLM KV-Cache Management:** Exploring how cost-aware caching principles can guide the retention, eviction, and placement of reusable prefix KV states across GPU and host memory
- **LLM Serving Systems:** Investigating the impact of KV-cache policies on latency, throughput, and memory efficiency in long-context, multi-turn, and agentic workloads

Education

Pohang University of Science and Technology (POSTECH) Mar 2024 – Present
Dept. of Computer Science and Engineering
M.S./Ph.D. Integrated Program
Data Systems Lab
Advisor: Prof. Wook-Shin Han

POSTECH, Dept. of Computer Science and Engineering Feb 2020 – Feb 2024
B.S. in Computer Science and Engineering
GPA: 4.07 / 4.3
Graduated Summa Cum Laude

Professional Experiences

Data Systems Lab, POSTECH, South Korea Jun 2022 – Feb 2024
Research Intern

- Fixed recovery correctness bugs in a state-of-the-art parallel logging scheme for in-memory DBMSs
- Fixed and optimized a competing parallel recovery scheme, improving its throughput by up to 2x
- Implemented operation-level workload capture and replay for database replay research
- Implemented efficient data loading for a distributed database system

Publications

1. **Taejun Yoon**[*], Taesung Lee[*], Joo Yeon Lee, Ki Hong Kim, Byungchul Tak, Wook-Shin Han. ([*]equal contribution) *Cache Replacement in Direct-Access Cloud OLAP—An Experimental Comparison and Practical Guidance.* Proc. ACM Manag. Data (SIGMOD 2027), to appear.

Projects

Prefix KV Cache Replacement and Tiering for LLM Serving Jul 2026 – Present

- **Problem:** In long-context and agentic LLM serving, which prefix KV blocks are retained and whether they reside in GPU or host memory largely determine prefill latency, yet existing replacement policies each hard-code a single reuse assumption and are brittle across workloads
- **Contribution:** Designed an online replacement and tiering policy that estimates each block's reuse hazard with a stratified recurrent-event model and derives both eviction and GPU-memory placement from a single Gittins index, recovering recency, frequency, and lifespan-based retention as special cases

Improving Sparse Attention for Long-Context LLM Inference Jul 2026 – Present

- **Problem:** Long-context decoding is bottlenecked by the GPU memory required for the KV cache. RetroInfer (Microsoft Research, VLDB 2026) offloads the KV cache to host memory and retrieves only the relevant tokens, but its fixed retrieval budgets trade accuracy for efficiency
- **Contribution:** Implemented RetroInfer on vLLM, optimized its serving throughput, and characterized its accuracy on agentic coding workloads (Terminal-Bench); currently extending RetroInfer with adaptive retrieval budgets to further improve the accuracy–efficiency trade-off

Semantic Query Processing with LLM-Powered Operators May 2026 – Present

- **Problem:** LLM-powered semantic operators enable queries over unstructured data, but each operator has many implementations with different quality, latency, and cost
- **Contribution:** Collaborating on optimizing and executing pipelines of semantic operators

Cache Replacement for Direct-Access Cloud OLAP Mar 2024 – Jun 2026

Industry Collaboration with SAP Labs Korea

- **Problem:** Direct-access cloud OLAP engines rely on local caches to hide remote object-store latency, but existing systems largely inherit recency-oriented policies from DBMS buffer management, ignoring that cloud misses combine a fixed per-request overhead with a size-dependent transfer cost
- **Contribution:** Conducted the first broad experimental comparison of 22 cache replacement algorithms for direct-access cloud OLAP, formulated a latency-aware metric (Unified Hit Ratio) validated against live S3 replay, derived practical policy-selection guidance, and instantiated it in a compact reference policy
- **Evaluation:** Collected workload traces from live executions of the commercial SAP HANA system over Parquet data on S3, grounding the guidance in a production system

Technical Skills

Programming Languages	C/C++ (Advanced), Python (Advanced), SQL (Advanced)
Data Systems	DuckDB, Apache Spark
Systems & Build Tools	Linux, Git, Docker, CMake
Cloud	AWS (EC2, S3), Azure (Virtual Machines, Blob Storage)
ML/LLM	PyTorch, vLLM, CUDA, OpenAI API
Languages	Korean (Native), English (Professional)

Honors and Fellowships

- Bang Sung-Yang Graduate Fellowship, Dept. of Computer Science and Engineering, POSTECH, 2024

Teaching Experiences

- **Teaching Assistant**, CSED321: Programming Languages, POSTECH, Spring 2026
- **Teaching Assistant**, CSED421: Database Systems, POSTECH, Spring 2024

Service

Lab Server Administrator, Data Systems Lab, POSTECH May 2024 – Present

- Built a web-based reservation system and server-down monitoring for the lab's shared GPU servers

References

Wook-Shin Han wshan@dblab.postech.ac.kr
Distinguished Professor, POSTECH, South Korea

Byungchul Tak bctak@knu.ac.kr
Associate Professor, School of Computer Science and Engineering, Kyungpook National University, South Korea