Yihao Sun

ysun67@syr.edu 3154508727

I am a 5^{th} year PhD researcher specialize in Database, HPC, and Programming Languages, focusing on massively parallel deductive database systems. I have published four papers in top conferences—**ASPLOS, NeurIPS, CLUSTER, VLDB**. My systems scale efficiently on ANLF's Theta and Polaris supercomputers.

Skill

- Proficient in programming languages, including C++, Python, Racket, Haskell, and Rust.
- Skilled in GPGPU programming, including CUDA/HIP kernel development and parallel libraries such as Thrust, CUB, and Kokkos.Experienced in GPGPU Programming. Familiar with CUDA/HIP programming writing CUDA kernel, parallel library Thrust/CUB/Kokkos.
- Experienced in scaling MPI programs on high-performance heterogenous clusters and supercomputers with accelerators such as Intel KNL and NVIDIA GPUs.
- Familiar with distributed programming and big data tools such as Apache Spark, Arrow, and Parquet. Experienced in techniques like data partitioning, caching, and load balancing.
- Knowledgeable in OLAP databases, particularly column-oriented and in-memory systems, with experience implementing a Datalog engine on DuckDB and optimizing large distributed queries.
- Skilled in large-scale graph query, with a focus on distributed and parallel environment.
- Knowledgeable in static program analysis, particularly in CFA and abstract interpretation.

Education

2020 – Present **Ph.D., CS, Syracuse University (Advised by Kristopher Micinski)**

2018 – 2020 **M.S, Computer Science, Syracuse University**

2013 – 2017 **B.S. in Electronic Engineer, Hangzhou Dianzi University**

Research Publications

Peer-Reviewed Papers

1. Y. Sun, S. Ahmedur, G. Thomas, M. Kristopher, and K. Sidharth, "Optimizing Datalog for the GPU," in *International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS)*, Rotterdam, The Netherlands, 2025. URL: https://arxiv.org/abs/2311.02206/.
2. G. Thomas, S. Arash, S. Yihao, K. Sowmith, K. Sidharth, and M. Kristopher, "Datalog with First-Class Facts," London, United Kingdom, 2025. URL: https://arxiv.org/pdf/2211.11573.
3. C. Liu, R. Saul, Y. Sun, *et al.*, "Assemblage: Automatic Binary Dataset Construction for Machine Learning," in *NeurIPS 2024 Datasets and Benchmarks Track*, 2024. URL: https://arxiv.org/pdf/2405.03991.
4. Y. Sun, K. Sidharth, G. Thomas, and K. Micinski, "Communication-Avoiding Recursive Aggregation," in *IEEE International Conference on Cluster Computing (CLUSTER)*, Santa Fe, USA, 2023, pp. 197–208. URL: https://kmicinski.com/assets/cluster23.pdf.
5. Y. Sun, J. Ching, and K. Micinski, "Declarative Demand-Driven Reverse Engineering," in *NDSS Workshop on Binary Analysis Research (BAR)*, Online, 2021. URL: https://arxiv.org/abs/2101.04718.
6. Y. Sun, K. Sidharth, G. Thomas, and K. Micinski, "So You Want To Analyze Scheme Programs With Datalog?" In *International Conference on Functional Programming(ICFP) Scheme Workshop*, Online, 2021. URL: https://arxiv.org/abs/2107.12909.

In Submission

1 "Column-Oriented Datalog on the GPU," *submitted to top conference in related area, under reviewing, but get very good scores (7 8 7 5).*, 2025.

Employment History

2017 – 2018 **Technical support engineer** Dahua Technology in China/India

2019-2020 **Teaching Assistant** Syracuse Univeristy, CIS352 Principle of Programming Language

2020-now **Research Assistant** Syracuse Univeristy, funded by Laboratory For Physical Science/V-SPELL Project.

Code Project

2021 **Slog**, MPI-based dialect of datalog specializing in parallel structural deduction reasoning. Implementing recursive aggregation and successfully scaling it to Thetha supercomputer in ANLF using a novel data partition technique. Solved a scaling bottleneck by dynamic load balancing and join order optimization, successful scaling system to 512 nodes. `https://github.com/harp-lab/slog-lang1`

2022 **Assemblage**, Developed a system capable of automated scraping and building from GitHub, collecting binaries for machine learning-based binary analysis. `https://huggingface.co/papers/2405.03991`

2023 **GDlog/VFlog**, a column-oriented GPU-based Datalog deductive database backend, outperforming state-of-the-art datalog engine Soufflé by 40× in numerous recursive query test cases. Old artifact for paper review `https://github.com/harp-lab/gdlog`. New artifact under developing `https://github.com/harp-lab/vflog`.

2024 **Ascent++**, enhanced version of Ascent, a Rust macro base datalog engine. Support restricted non-stratified negation and functional programming. `https://github.com/StarGazerM/ascent-plusplus`