Lin Ma

Carnegie Mellon University
Department of Computer Science
Gates-Hillman Center 4111
Pittsburgh, PA 15213-3891 USA

Voice: +1-412-519-7097
E-mail: lin.ma@cs.cmu.edu
Web: http://www.cs.cmu.edu/~malin199
GitHub: https://github.com/malin1993ml

Current Position

Postdoctoral Researcher 2021-present
Carnegie Mellon University *Pittsburgh, PA USA*
Supervisor: Andy Pavlo

Education

Ph.D., Computer Science 2015-2021
Carnegie Mellon University *Pittsburgh, PA USA*
Advisor: Andy Pavlo

M.Sc., Computer Science 2015-2018
Carnegie Mellon University *Pittsburgh, PA USA*

B.Sc., Computer Science 2011-2015
Peking University *Beijing, China*
Advisor: Bin Cui

Research Experience

Graduate Research Assistant 2015-2021
Carnegie Mellon University

- **NoisePage** - https://noise.page/
  A new in-memory relational DBMS designed from group up to manage itself autonomously and remove the need for human administration. My research focuses on the following areas:

  Self-Driving DBMS: System architecture inspired by self-driving vehicles to enable autonomous operations with three components: workload forecasting, behavior modeling, and action planning.

  Workload Forecasting: Framework that predicts trends and patterns of the queries in the workload using query templatization, arrival-rate clustering, and an ensemble of ML models.

  Behavior Modeling: Framework that models the impact of self-driving actions (e.g., creating indexes) by decomposing the system into small and independent tasks to build models separately.

  Action Planning: Framework that uses receding horizon planning and Monte Carlo tree search to plan for action sequences to apply given the forecasted workload and estimated action behavior.

- **H-Store** - http://hstore.cs.brown.edu

  Modern Storage Hardware: Evaluation of the design decisions on managing cold data in in-memory DBMSs tailored to different storage devices, including NVMs, SSDs, HDDs, and SMRs.

Research Intern Summer 2018
Data Management, Exploration and Mining Group
Microsoft Research, Redmond

- **Auto-Indexing in Cloud** - https://www.microsoft.com/en-us/research/project/autoadmin

  Holistic Active Learner: Data collection mechanism that leverages B-instances in the cloud and active learning to acquire additional labels to improve the models used by ML enhanced DBMSs.

Undergraduate Research Intern 2013-2015
Peking University

- **Graph Computation**

  Social Network Analysis: System prototype built from scratch to answer four types of social queries, including finding large interest communities and the most central people.

  PSgL: Parallel subgraph listing framework built on top of Apache Giraph that iteratively divides the problem into partial tasks and balances the loads between concurrent workers.

  Page: Graph computation engine built on top of Apache Giraph that uses the online graph partitioning statistics to guide the resource allocation for parallel processing.

Publications

[1] Matthew Butrovich, Wan Shen Lim, **Lin Ma**, John Rollinson, William Zhang, Yu Xia, and Andrew Pavlo. Tastes great! less filling! high performance and accurate training data collection for self-driving database management systems. In *Proceedings of the 2022 ACM SIGMOD International Conference on Management of Data*, 2022.

[2] **Lin Ma**, William Zhang, Jie Jiao, Wuwen Wang, Matthew Butrovich, Wan Shen Lim, Prashanth Menon, and Andrew Pavlo. Mb2: Decomposed behavior modeling for self-driving database management systems. In *Proceedings of the 2021 ACM SIGMOD International Conference on Management of Data*, pages 1248–1261, 2021.

[3] Andrew Pavlo, Matthew Butrovich, **Lin Ma**, Prashanth Menon, Wan Shen Lim, Dana Van Aken, and William Zhang. Make your database system dream of electric sheep: Towards self-driving operation. *Proceedings of the VLDB Endowment*, 14(12):3211–3221, 2021.

[4] Amadou Ngom, Prashanth Menon, Matthew Butrovich, **Lin Ma**, Wan Shen Lim, Todd C Mowry, and Andrew Pavlo. Filter representation in vectorized query execution. In *Proceedings of the 17th International Workshop on Data Management on New Hardware (DaMoN 2021)*, pages 1–7, 2021.

[5] Ling Zhang, Matthew Butrovich, Tianyu Li, Andrew Pavlo, Yash Nannapaneni, John Rollinson, Huanchen Zhang, Ambarish Balakumar, Daniel Biales, Ziqi Dong, Emmanuel J. Eppinger, Jordi E. Gonzalez, Wan Shen Lim, Jianqiao Liu, **Lin Ma**, Prashanth Menon, Soumil Mukherjee, Tanuj Nayak, Amadou Ngom, Dong Niu, Deepayan Patra, Poojita Raj, Stephanie Wang, Wuwen Wang, Yao Yu, and William Zhang. Everything is a transaction: Unifying logical concurrency control and physical data structure maintenance in database management systems. In *11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings*, 2021.

[6] **Lin Ma**, Bailu Ding, Sudipto Das, and Adith Swaminathan. Active learning for ml enhanced database systems. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 175–191, 2020.

[7] Prashanth Menon, Amadou Ngom, **Lin Ma**, Todd C Mowry, and Andrew Pavlo. Permutable compiled queries: dynamically adapting compiled queries without recompiling. *Proceedings of the VLDB Endowment*, 14(2):101–113, 2020.

[8] Andrew Pavlo, Matthew Butrovich, Ananya Joshi, **Lin Ma**, Prashanth Menon, Dana Van Aken, Lisa Lee, and Ruslan Salakhutdinov. External vs. internal: an essay on machine learning agents for autonomous database management systems. *IEEE bulletin*, 42(2), 2019.

[9] **Lin Ma**, Dana Van Aken, Ahmed Hefny, Gustavo Mezerhane, Andrew Pavlo, and Geoffrey J. Gordon. Query-based workload forecasting for self-driving database management systems. In *Proceedings of the 2018 ACM International Conference on Management of Data*, pages 631–645, 2018.

[10] Andrew Pavlo, Gustavo Angulo, Joy Arulraj, Haibin Lin, Jiexi Lin, **Lin Ma**, Prashanth Menon, Todd C Mowry, Matthew Perron, Ian Quah, et al. Self-driving database management systems. In *CIDR*, 2017.

[11] **Lin Ma**, Joy Arulraj, Sam Zhao, Andrew Pavlo, Subramanya R Dulloor, Michael J Giardino, Jeff Parkhurst, Jason L Gardner, Kshitij Doshi, and Stanley Zdonik. Larger-than-memory data management on modern storage hardware for in-memory oltp database systems. In *Proceedings of the 12th International Workshop on Data Management on New Hardware*, page 9. ACM, 2016.

[12] Huanchen Zhang, David G Andersen, Andrew Pavlo, Michael Kaminsky, **Lin Ma**, and Rui Shen. Reducing the storage overhead of main-memory oltp databases with hybrid indexes. In *Proceedings of the 2016 International Conference on Management of Data*, pages 1567–1581. ACM, 2016.

[13] Yingxia Shao, Bin Cui, and **Lin Ma**. Page: a partition aware engine for parallel graph computation. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):518–530, 2015.

[14] Yingxia Shao, Bin Cui, Lei Chen, **Lin Ma**, Junjie Yao, and Ning Xu. Parallel subgraph listing in a large-scale graph. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 625–636. ACM, 2014.

[15] Yingxia Shao, Junjie Yao, Bin Cui, and **Lin Ma**. Page: A partition aware graph computation engine. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 823–828. ACM, 2013.

Teaching

- **Instructor** - 15-445/645 Introduction to Database Systems
  Carnegie Mellon University, 2021
  Delivered half of the course lectures, including topics in query optimization, concurrency control, logging and recovery, and distributed databases. Managed several TAs and more than 100 enrolled students with the other instructor. Planned homework, projects, and exams with help from TAs.
- **Head Teaching Assistant** - 15-721 Advanced Database Systems
  Carnegie Mellon University, 2021
- **Teaching Assistant** - 15-445/645 Introduction to Database Systems
  Carnegie Mellon University, 2021

Awards and Scholarships

- The China Computer Federation Outstanding Undergraduate Award - 2015
- China National Scholarship - 2014
- SIGMOD Programming Contest Finalist - 2014
- SIGMOD Travel Award - 2014

Service

To the Profession

- Web/Information Chair and Program Committee - SIGMOD 2023
- Program Committee - VLDB 2022
- Program Committee - SMDB@ICDE 2022
- Program Committee - AIDB@VLDB 2021
- Program Committee - AIDB@VLDB 2020
- External Reviewer - DAPD 2019
- External Reviewer - SIGMOD Demo 2017

To the University

- CSD Faculty Search Committee – Carnegie Mellon University, 2020
- CSD MS Admissions Committee – Carnegie Mellon University, 2018
- Graduate Student Recruitment (Open House) Committee – Carnegie Mellon University, 2018

Academic Talks

- **NoisePage: The Self-Driving Database Management System**
  Ahana, October 19, 2021

University of California, San Diego, October 6, 2021
Facebook, June 4, 2021
Harvard University, April 30, 2021
Columbia University, April 13, 2021
Stanford University (MLSys Seminar), April 8, 2021
Oracle, April 6, 2021
Carnegie Mellon University, March 22, 2021
Centrum Wiskunde & Informatica, March 19, 2021
The University of Chicago, March 17, 2021
University of Washington, March 3, 2021
University of California, Berkeley, February 23, 2021
University of California, Santa Cruz (CSE 215), February 19, 2021
Technical University of Munich, February 18, 2021
Brown University, January 27, 2021

- **MB2: Decomposed Behavior Modeling for Self-Driving Database Management Systems**
  SIGMOD, June 2021
- **Active Learning for ML Enhanced Database Systems**
  SIGMOD, June 2020
- **Self-Driving Databases: It All Starts with Workload Forecasting**
  Percona Live, May 2019
- **Efficiently Leveraging B-Instances for Query Plan Predictions**
  Microsoft Research, August 2018
- **Query-based Workload Forecasting for Self-Driving DBMSs**
  SIGMOD, June 2018
  Microsoft Research, May 2018
  PDL Retreat, October 2017
- **Larger-than-Memory Data Management on Modern Storage Hardware for In-Memory OLTP Database Systems**
  SIGMOD, June 2016
- **The Self-Driving DBMS**
  PDL Retreat, October 2016
- **Multi-Level Anti-Caching for NVM+SSD in H-Store**
  PDL Retreat, October 2015
- **Finalist Presentation of Programming Contest**
  SIGMOD, June 2014
- **Using Less to Do More With Anti-Caching in OLTP Database Systems**
  Carnegie Mellon University, August 2014