Huize Li

+1-407-590-0278 | huize.li@ucf.edu | Personal Homepage

in LinkedIn Page | ⬤ Github | ⬤ Google Scholar | ⓇG ResearchGate | ⒾD ORCID

Orlando, FL - 32825, USA

RESEARCH INTERESTS

I am currently working on machine learning accelerators, Transformer and sparse attention, in-memory computing, in-situ computing, and domain specific accelerators. I am also desired to explore photonic computing and hyper-dimensional computing.

EDUCATION

• **Huazhong University of Science and Technology [⬤]**   *Sept 2017 - Dec 2022*
Ph.D. in Engineering   Wuhan, China
  ◦ Major: Computer Architecture
  ◦ Mentors: Professor Hai Jin, IEEE Fellow (email: hjin@hust.edu.cn).

• **Huazhong University of Science and Technology [⬤]**   *Sept 2013 - Jun 2017*
B.S. in Engineering   Wuhan, China
  ◦ Major: Software Engineering.
  ◦ GPA: 3.5/4.0, School number: 027-87541114.

TEACHING EXPERIENCE

• **Huazhong University of Science and Technology [⬤]**   *Fall 2017*
Teaching Assistant, Graduate Programs: Parallel Processing   Wuhan, China
  ◦ Duty: Preparing, guiding, and evaluating semester research projects.

• **Huazhong University of Science and Technology [⬤]**   *Fall 2019*
Teaching Assistant, Graduate Programs: Advanced Computer Architecture   Wuhan, China
  ◦ Duty: Preparing, guiding, and evaluating semester research projects.

EMPLOYMENT

• **Department of Electrical and Computer Engineering in University of Central Florida [⬤]** *Feb 2025 - ongoing*
Postdoctoral Researcher   Orlando, FL, USA
  ◦ Mentors: Professor Xin Xin (email: xin.xin@ucf.edu).
  ◦ Duty: I do researches and projects in memory reliability and processing in memory architecture.

• **School of Computing in National University of Singapore [⬤]**   *Feb 2023 - Feb 2025*
Postdoctoral Research Fellow   Singapore
  ◦ Mentors: Professor Tulika Mitra (email: tulika@comp.nus.edu.sg).
  ◦ Duty: I do researches and projects in design high performance and
    energy-efficient accelerator for sparse Transformer.

PUBLICATIONS   C=CONFERENCE, J=JOURNAL, P=PATENT, S=IN SUBMISSION, T=THESIS

[T.1] <u>**Huize Li**</u>. (2022). Processing-in-Memory Architecture Based Structured Query Accelerators. Ph.D. Thesis.

[C.7] Chaoqiang Liu, Dan Chen, Yu Huang, Wenjing Xiao, Haifeng Liu, Yi Zhang, **Huize Li**, Xiaofei Liao, and Hai Jin. (2025). SeIM: In-Memory Acceleration for Approximate Nearest Neighbor Search. In *Proceedings of the 62th ACM/IEEE Design Automation Conference (DAC)*, Just Accepted.

[C.6] Zhaoying Li, Dan Wu, Dhananjaya Wijerathne, Dan Chen, **Huize Li**, Cheng Tan, and Tulika Mitra. (2025). Rewire: Advancing CGRA Mapping Through a Consolidated Routing Paradigm. In *Proceedings of the 62th ACM/IEEE Design Automation Conference (DAC)*, Just Accepted.

[C.5] **Huize Li**, Dan Chen, and Tulika Mitra. (2025). HyAtten: Hybrid Photonic-digital Architecture for Accelerating Attention Mechanism. *Design, Automation, and Test in Europe (DATE)*, Just Accepted.

[C.4] **Huize Li**, Zhaoying Li, Zhenyu Bai, and Tulika Mitra. (2024). ASADI: Accelerating Sparse Attention using Diagonal-based In-situ Computing. In *Proceedings of the 30th IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 774-787.

[C.3] Zhenyu Bai, Pranav Dangi, **Huize Li**, and Tulika Mitra. (2024). SWAT: Scalable and Efficient Window Attention-based Transformers Acceleration on FPGAs. In *Proceedings of the 61th ACM/IEEE Design Automation Conference (DAC)*, pp. 93:1-93:6.

[C.2] <u>**Huize Li**</u>, Hai Jin, Long Zheng, Yu Huang, Xiaofei Liao, Zhuohui Duan, Dan Chen, and Chuangyi Gui. (2022). ReSMA: accelerating approximate string matching using ReRAM-based content addressable memory. In *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*, pp. 991-996.

[C.1] Cong Liu, Haikun Liu, Hai Jin, Xiaofei Liao, Yu Zhang, Zhuohui Duan, Jiahong Xu, and <u>**Huize Li**</u>. (2022). ReGNN: a ReRAM-based heterogeneous architecture for general graph neural networks. In *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*, pp. 469-474.

[J.7] <u>**Huize Li**</u>, Dan Chen, and Tulika Mitra. (2025). SPLIM: Bridging the Gap Between Unstructured SpGEMM and Structured In-situ Computing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (IEEE TCAD)*, Just Accepted.

[J.6] <u>**Huize Li**</u>, Dan Chen, and Tulika Mitra. (2025). SADIMM: Accelerating Sparse Attention using DIMM-based Near-memory Processing. *IEEE Transactions on Computers (IEEE TC)*, 74 (2), pp. 542-554.

[J.5] Cong Liu, Kaibo Wu, Haikun Liu, Hai Jin, Xiaofei Liao, Zhuohui Duan, Jiahong Xu, <u>**Huize Li**</u>, Yu Zhang, and Jing Yang. (2025). A ReRAM-Based Processing-In-Memory Architecture for Hyperdimensional Computing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (IEEE TCAD)*, 44 (2), pp. 512-524.

[J.4] <u>**Huize Li**</u>, Hai Jin, Long Zheng, Yu Huang, Xiaofei Liao, Dan Chen, Zhuohui Duan, Cong Liu, Jiahong Xu, and Chuanyi Gui. (2024). CPSAA: Accelerating Sparse Attention using Crossbar-based Processing-In-Memory Architecture. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (IEEE TCAD)*, 43 (6), pp. 1741-1754.

[J.3] Jiahong Xu, Haikun Liu, Zhuohui Duan, Xiaofei Liao, Hai Jin, Xiaokang Yang, <u>**Huize Li**</u>, Cong Liu, Fubing Mao, and Yu Zhang. (2024). ReHarvest: an ADC Resource-Harvesting Crossbar Architecture for ReRAM-Based DNN Accelerators. *ACM Trans. Archit. Code Optim. (ACM TACO)*, 21 (3), pp. 1-26.

[J.2] <u>**Huize Li**</u>, Hai Jin, Long Zheng, Yu Huang, and Xiaofei Liao. (2022). ReCSA: a dedicated sort accelerator using ReRAM-based content addressable memory. *Frontiers of Computer Science (FCS)*, 17: 172103.

[J.1] <u>**Huize Li**</u>, Hai Jin, Long Zheng, and Xiaofei Liao. (2020). ReSQM: Accelerating Database Operations Using ReRAM-Based Content Addressable Memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (IEEE TCAD)*, 39 (11), pp. 4030-4041.

PROFESSIONAL SERVICES

Reviewer
Reviewer of IEEE TC.

Services

TALKS

ASADI: Accelerating Sparse Attention using Diagonal-based In-situ Computing. *HPCA 2024*.
ReSMA: accelerating approximate string matching using ReRAM-based content addressable memory. *DAC 2022*.
ReSQM: Accelerating Database Operations Using ReRAM-Based Content Addressable Memory. *CODES+ISSS 2020*.

REFERENCES

1. **Hai Jin**
   Professor, Department of Computer Science
   Huazhong University of Science and Technology
   Email: hjin@hust.edu.cn
   Relationship: [Ph.D. Advisor]

2. **Tulika Mitra**
   Professor, School of Computing
   National University of Singapore
   Email: tulika@comp.nus.edu.sg
   Relationship: [My Postdoc Mentor]