

Huai, Shuo Ph.D. Candidate

✉ shuohuai@outlook.com

🐙 github.com/shuo-huai

🎓 @Google Scholar

🌐 www.linkedin.com/in/shuo-huai-7351001a0/

🌐 shuo-huai.github.io



I received the BSc degree from the School of Computer Science and Technology, Shandong University, Jinan, China, in 2019. I am currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. My research interests are embedded intelligence, neural network optimization, in-memory computing, and non-volatile memory.

Education

- | | | |
|-----------------------|---|--------------------------------------|
| Jan. 2020 – Present | 🎓 Ph.D., Nanyang Technological University
Ph.D. in Computer Science and Engineering | <i>Supervisor: Prof. Liu Weichen</i> |
| Sep. 2015 – Jun. 2019 | 🎓 BSc, Shandong University
Bachelor of Computer Science and Technology | <i>Advisor: Prof. Zhao Mengying</i> |

Research Publications

Conference Proceedings

- 1 H. Chen, D. Liu, S. Li, **S. Huai**, X. Luo, and W. Liu, “Mugnoc: A software-configured multicast-unicast-gather noc for accelerating cnn dataflows,” in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, 2023, pp. 308–313.
- 2 **S. Huai**, D. Liu, X. Luo, H. Chen, W. Liu, and R. Subramaniam, “Crossbar-aligned & integer-only neural network compression for efficient in-memory acceleration,” in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, 2023, pp. 234–239.
- 3 H. Kong, X. Luo, **S. Huai**, *et al.*, “Emnape: Efficient multi-dimensional neural architecture pruning for edgeai,” in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2023, pp. 1–2.
- 4 **S. Huai**, D. Liu, H. Kong, *et al.*, “Collate: Collaborative neural network learning for latency-critical edge systems,” in *2022 IEEE 40th International Conference on Computer Design (ICCD)*, IEEE, 2022, pp. 627–634.
- 5 H. Kong, D. Liu, **S. Huai**, *et al.*, “Smart scissor: Coupling spatial redundancy reduction and cnn compression for embedded hardware,” in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.
- 6 X. Luo, D. Liu, H. Kong, **S. Huai**, H. Chen, and W. Liu, “Work-in-progress: What to expect of early training statistics? an investigation on hardware-aware neural architecture search,” in *2022 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ ISSS)*, IEEE, 2022, pp. 1–2.
- 7 X. Luo, D. Liu, H. Kong, **S. Huai**, H. Chen, and W. Liu, “You only search once: On lightweight differentiable architecture search for resource-constrained embedded platforms,” in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 475–480.
- 8 **S. Huai**, L. Zhang, D. Liu, W. Liu, and R. Subramaniam, “Zerobn: Learning compact neural networks for latency-critical edge systems,” in *2021 58th ACM/IEEE Design Automation Conference (DAC)*, IEEE, 2021, pp. 151–156.
- 9 X. Luo, D. Liu, **S. Huai**, and W. Liu, “Hsconas: Hardware-software co-design of efficient dnns via neural architecture search,” in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2021, pp. 418–421.
- 10 **S. Huai**, W. Song, M. Zhao, X. Cai, and Z. Jia, “Performance-aware wear leveling for block ram in nonvolatile fpgas,” in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.


Journal Articles

- 1 **S. Huai**, H. Kong, S. Li, *et al.*, “Evolp: Self-evolving latency predictor for model compression in real-time edge systems,” *IEEE Embedded Systems Letters*, 2023.



- 2 **S. Huai**, H. Kong, X. Luo, *et al.*, "Crimp: Compact & reliable dnns inference for in-memory processing via crossbar-aligned compression and non-ideality adaptation," *ACM Transactions on Embedded Computing Systems*, 2023.
- 3 **S. Huai**, H. Kong, X. Luo, *et al.*, "On hardware-aware design and optimization of edge intelligence," *IEEE Design & Test*, 2023.
- 4 **S. Huai**, D. Liu, H. Kong, *et al.*, "Latency-constrained dnn architecture learning for edge systems using zerorized batch normalization," *Future Generation Computer Systems*, vol. 142, pp. 314–327, 2023.
- 5 H. Kong, D. Liu, **S. Huai**, *et al.*, "Edgecompress: Coupling multi-dimensional model compression and dynamic inference for edgeai," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.
- 6 S. Li, **S. Huai**, and W. Liu, "An efficient gustavson-based sparse matrix-matrix multiplication accelerator on embedded fpgas," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.
- 7 H. Chen, P. Chen, X. Luo, **S. Huai**, and W. Liu, "Lamp: Load-balanced multipath parallel transmission in point-to-point nocs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 12, pp. 5232–5245, 2022.
- 8 X. Luo, D. Liu, H. Kong, **S. Huai**, H. Chen, and W. Liu, "Lightnas: On lightweight and scalable neural architecture search for embedded platforms," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2022.
- 9 X. Luo, D. Liu, H. Kong, **S. Huai**, H. Chen, and W. Liu, "Surgenas: A comprehensive surgery on hardware-aware differentiable neural architecture search," *IEEE Transactions on Computers*, vol. 72, no. 4, pp. 1081–1094, 2022.
- 10 H. Kong, **S. Huai**, D. Liu, *et al.*, "Edlab: A benchmark for edge deep learning accelerators," *IEEE Design & Test*, vol. 39, no. 3, pp. 8–17, 2021.
- 11 X. Luo, D. Liu, **S. Huai**, H. Kong, H. Chen, and W. Liu, "Designing efficient dnns via hardware-aware neural architecture search and beyond," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 6, pp. 1799–1812, 2021.

Experience





Work

- Sep. 2019 – Oct. 2023  **Project Officer, HP-NTU Digital Manufacturing Corporate Lab**
 Machine Learning Optimization for Edge Devices
 – Design a performance evaluation methodology and toolset for Machine Learning Accelerators.
 – Provide HP business groups, like Business Personal Systems and Digital Manufacturing, the insights, and framework for process and resource-efficient design and optimization of DNN, accelerators, and system architectures for machine learning/inference at the edge.

Teaching

- Feb. 2020 – Mar. 2020  **Teaching Assistant, Nanyang Technological University**
 CE/CZ 1006 Computer Organisation & Architecture
- Feb. 2019 – Jun. 2019  **Teaching Assistant, Shandong University**
 SD01331470 Computer Organization and Design

Awards and Achievements

- 2023  **Support Grant**, ACM SIGDA Student Research Forum, ASP-DAC 2023, Japan.
- 2019  **Outstanding Undergraduate Graduation Dissertation**, Shandong University.
- 2018  DAC system Design Contest (6/61).
- 2017  **Meritorious Winner**, COMAP's Mathematical Contest In Modeling.

Innovations

Patent

- **A Non-Volatile FPGA Placement Optimization Method and System Based on Performance-aware Wear Leveling**

Inventors: 1) Zhao Mengying; 2) **Huai Shuo**; 3) Shen Zhaoyan; 4) Cai Xiaojun; 5) Jia Zhiping

Filed date: 20 May 2019

Patent No.: ZL 2019 1 0419760.9

Technology Disclosure (TD)

- **EDLAB: A Benchmark Tool for Edge Deep Learning Accelerators**

Inventors: 1) Liu Weichen; 2) Liu Di; 3) Kong Hao; 4) Zhang Lei; 5) **Huai Shuo**; 6) Li Shiqing; 7) Chen Hui; 8) Zhu Shien

Filed date: 08 July 2020

Ref: TD 2020-264

- **ZeroBN: Learning Compact Neural Networks For Latency-Critical Edge Systems**

Inventors: 1) Liu Weichen; 2) **Huai Shuo**; 3) Liu Di; 4) Zhang Lei

Filed date: 09 March 2021

Ref: TD 2021-096

- **Collaborative Neural Network Learning For Multiple Latency-Critical Edge Systems**

Inventors: 1) Liu Weichen; 2) **Huai Shuo**; 3) Kong Hao

Filed date: 15 August 2022

Ref: TD 2022-287

- **Smart Scissor: A Deep Compression Framework For Jointly Reducing the Redundancy In Images And Neural Networks**

Inventors: 1) Liu Weichen; 2) Kong Hao; 3) **Huai Shuo**

Filed date: 15 August 2022

Ref: TD 2022-288

Skills

Languages	■ Professional working proficiency in English, and native proficiency in Mandarin Chinese.
Coding & OS	■ Python, Java, C/C++, Linux, SQL, \LaTeX , and Verilog