HANCHEN YE

403 Coordinated Science Lab, 1308 W Main St, Urbana, IL 61801, USA
Email: hanchenye@gmail.com ⋄ Homepage: https://hanchenye.com

EDUCATION

- **University of Illinois at Urbana-Champaign**, Urbana, IL
  Ph.D. in Electrical and Computer Engineering — *Aug. 2019 - Present*
  Thesis Advisor: Prof. Deming Chen
- **Fudan University**, Shanghai, China
  M.E. in Integrated Circuit Engineering — *Sep. 2017 - Jun. 2019*
  B.E. in Microelectronic Science and Engineering — *Sep. 2013 - Jun. 2017*
- **National University of Singapore**, Singapore
  Exchange Student in Electrical and Computer Engineering — *Aug. 2015 - Dec. 2015*

WORK EXPERIENCES

- **University of Illinois at Urbana-Champaign**, Urbana, IL
  Research Assistant (Full-time) in Electrical and Computer Engineering — *Aug. 2019 - Present*
  Advisor: Prof. Deming Chen
  Teaching Assistant (Part-time) in ECE527 (System-On-Chip Design) — *Aug. 2023 - Dec. 2023*
  Teaching Assistant (Part-time) in ECE527 (System-On-Chip Design) — *Aug. 2022 - Dec. 2022*
- **Inspirit IOT**, Champaign, IL
  Part-time Intern (Part-time) — *Jan. 2024 - Dec. 2024*
  Mentor: Prof. Deming Chen
- **Google**, Mountain View, CA
  Ph.D. Resident (Internship) in X, The Moonshot Factory — *May. 2023 - Aug. 2023*
  Mentor: Xiaoqing Xu, Prof. David Pan, Chris Leary
- **Intel**, Portland, OR
  Research Intern (Internship) in Strategic CAD Labs — *May. 2022 - Aug. 2022*
  Mentor: Jin Yang, Jeremy Casas, Zhenkun Yang
- **SiFive**, San Mateo, CA
  Compilers Intern (Internship) in Platform Engineering Department — *May. 2021 - Aug. 2021*
  Mentor: Andrew Lenharth
- **Xilinx (AMD)**, San Jose, CA
  Compiler Intern (Internship) in Research Labs — *Jun. 2020 - Aug. 2020*
  Mentor: Stephen Neuendorffer
- **Fudan University**, Shanghai, China
  Research Assistant (Part-time) in State Key Laboratory of ASIC and System — *Sep. 2016 - Jun. 2019*
  Advisor: Prof. Gengsheng Chen

AWARDS AND SCHOLARSHIPS

- **UIUC** Dr. Ok Kyun Kim Fellowship — *Mar. 2024*
- **UIUC** Conference Presentation Awards — *Mar. 2024*
- **SRC TECHCON** First Place Best Student Presenter Award — *Sep. 2023*
- **DAC Ph.D. Forum** First Place Winner — *Jul. 2023*

- **UIUC** A.R. Buck Knight Fellowship *Apr. 2023*
- **AMD** HACC Outstanding Researcher Awards *Feb. 2023*
- **UIUC** Teachers Ranked as Excellent *Dec. 2022*
- **UIUC** Rambus Computer Engineering Fellowship *May. 2022*
- **DAC** Young Fellows *Jun. 2020, Apr. 2022*
- **Shanghai** Outstanding Graduates *Jun. 2019*
- **Fudan University** KLA-Tencor Scholarship *Dec. 2018*
- **Fudan University** Outstanding Graduate Students *Oct. 2018*
- **The 2nd China College IC Competition** Grand Prize Winner *Aug. 2018*
- **Fudan University** Xi-Yuan Research Scholarship *May. 2016*
- **Fudan University** Outstanding Undergraduate Students *Dec. 2015*

SELECTED PROJECTS

- **StreamTensor: Make Tensors Stream in Dataflow Accelerators** *Jan. 2024 - Present*
  Document: https://hanchenye.com/streamtensor
  - Designed a dataflow-centric typing system and intermediate representation (IR) to model the kernel processing and communication at tensor level in MLIR.
  - Introduced stream-based kernel fusion, on-the-fly memory layout conversion, and dataflow FIFO optimization to reduce off-chip memory access and on-chip memory size.
  - Designed a compilation pipeline that compiles PyTorch model, e.g., large language model (LLM), to low-level IRs targeting dataflow accelerators, such as AMD Versal ACAP and FPGA.
- **XLS: Accelerated HW Synthesis** *May. 2023 - Aug. 2023*
  Github: https://github.com/google/xls
  - XLS implements a High-level Synthesis (HLS) toolchain which produces synthesizable designs (Verilog and SystemVerilog) from flexible, high-level descriptions of functionality.
  - Proposed a feedback-directed optimization (FDO) method named ISDC that takes downstream tools, e.g., OpenROAD, results as feedback to improve SDC scheduling quality of HLS.
  - Achieved a 28.5
- **HIDA: A Hierarchical Dataflow Compiler for High-level Synthesis** *Mar. 2022 - Jan. 2024*
  Github: https://github.com/UIUC-ChenLab/ScaleHLS-HIDA
  - Proposed a hierarchical dataflow intermediate representation (IR) to model and optimize the complicated dataflow structures in High-level Synthesis (HLS).
  - Designed an algorithm to guide the local design space exploration of each dataflow node while keeping the global dataflow balanced and efficient.
- **CHARM: A Heterogeneous GEMM Accelerator on Versal ACAP** *Dec. 2021 - Oct. 2022*
  Github: https://github.com/arc-research-lab/CHARM
  - Mapped GEMM-based models, e.g., BERT and ViT, to accelerators on AMD Versal ACAP; Non-GEMM kernels and data movement kernels are implemented on Programming Logic (PL).
  - Proposed a design space exploration algorithm to determine the tiling strategy at each level of memory.
- **PolyAIE: A Polyhedral Compiler for Versal ACAP** *Oct. 2021 - Apr. 2022*
  Github: https://github.com/hanchenye/polyaie
  - Designed a compilation flow from C/C++ programs to the AI-Engine (AIE) array on AMD Versal ACAP using Polyhedral compilation techniques in MLIR.

- **CIRCT: Circuit IR Compilers and Tools** *Jun. 2020 - Oct. 2021*
  Github: https://github.com/llvm/circt
  - The CIRCT open-source project is an effort looking to apply MLIR and the LLVM development methodology to the domain of hardware design tools.
  - Contributed to the FIRRTL, HW (Hardware), and SV (SystemVerilog) dialects and transformations to establish the hardware 'core IR' of CIRCT and enable a Chisel to SystemVerilog compilation flow.
  - Contributed a new FSM dialect to represent, optimize, and generate codes for finite-state machines.
  - Contributed to the Handshake and Pipeline dialects to enable a High-level Synthesis (HLS) flow that compiles the 'core IR' of MLIR to the hardware 'core IR' of CIRCT.

- **ScaleHLS: A Scalable High-level Synthesis Framework on MLIR** *Apr. 2020 - Mar. 2022*
  Github: https://github.com/UIUC-ChenLab/scalehls
  - Designed a multi-level HLS representation and optimization framework in MLIR.
  - Designed an HLS-specific transform and analysis library, including loop and pragma optimizations, an HLS Quality-of-Result (QoR) estimator, and a multi-objective design space explorer.
  - Designed a C/C++ front-end and an HLS C/C++ code generator for MLIR.

- **DNNExplorer: A Novel Design Paradigm of DNN Accelerator** *Feb. 2020 - Mar. 2021*
  - Proposed a novel DNN acceleration paradigm which can take advantage of both dataflow pipeline and overlay architectures, enabling a more scalable solution compared to previous arts.
  - Proposed an efficient design space exploration algorithm to generate optimized DNN accelerators following the new paradigm.

- **HybridDNN: Hybrid Spatial and Winograd DNN Accelerator** *Jan. 2019 - Dec. 2019*
  - Proposed a hybrid Spatial and Winograd convolution architecture for DNN acceleration.
  - Designed a comprehensive tool for the performance and area estimation and the design space exploration for both edge and cloud FPGAs.

- **Musket: RISCV-based IoT Sensor-Hub on FPGA** *Apr. 2018 - Aug. 2018*
  - Pruned and transplanted a RISCV core to an edge FPGA and established a low-power SoC.
  - Ported an RTOS to manage sensors and the wireless connection between FPGA and smartphones.
  - Won the outstanding award of the 2nd China College IC Competition.

- **RS-Pipeline: Dynamic and Pipelined CNN Accelerator on FPGA** *Oct. 2017 - May. 2018*
  - Proposed a Dynamic Partial Reconfiguration (DPR) -based pipeline architecture to deploy large CNN accelerators on resource-limited FPGAs while maintaining a low overall latency.

PUBLICATIONS

[1] **[NeurIPS'24] SnapKV: LLM Knows What You are Looking for Before Generation**
Yuhong Li*, Yingbing Huang*, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, **Hanchen Ye**, Tianle Cai, Patrick Lewis, Deming Chen
The Conference on Neural Information Processing Systems (NeurIPS'24)

[2] **[TRETS] CHARM 2.0: Composing Heterogeneous Accelerators for Deep Learning on Versal ACAP Architecture (Journal)**
Jinming Zhuang, Jason Lau, **Hanchen Ye**, Zhuoping Yang, Shixin Ji, Jack Lo, Kristof Denolf, Stephen Neuendorffer, Alex Jones, Jingtong Hu, Yiyu Shi, Deming Chen, Jason Cong, Peipei Zhou
The ACM Transactions on Reconfigurable Technology and Systems (TRETS)

[3] **[LAD'24] An Iteratively-refined Dataset for High-Level Synthesis Functional Verification through LLM-Aided Bug Injection**
Lily Jiaxin Wan, **Hanchen Ye**, Jinghua Wang, Manvi Jha, Deming Chen
The IEEE International Workshop on LLM-Aided Design (LAD'24)

[4] **[DAC'24] New Solutions on LLM Acceleration, Optimization, and Application (Invited)**
Yingbing Huang, Lily Jiaxin Wan, **Hanchen Ye**, Manvi Jha, Jinghua Wang, Yuhong Li, Xiaofan Zhang,

Deming Chen
The ACM/IEEE Design Automation Conference (DAC'24)

[5] **[DATE'24] Subgraph Extraction-based Feedback-guided Iterative Scheduling for HLS**
Hanchen Ye, David Pan, Chris Leary, Deming Chen, Xiaoqing Xu
The Conference on Design, Automation & Test in Europe (DATE'24)

[6] **[ASPLOS'24] HIDA: A Hierarchical Dataflow Compiler for High-Level Synthesis**
Hanchen Ye, Hyegang Jun, Deming Chen
The ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'24)

[7] **[ASP-DAC'24] Software/Hardware Co-design for LLM and Its Application for Design Verification (Invited)**
Lily Jiaxin Wan*, Yingbing Huang*, Yuhong Li, **Hanchen Ye**, Jinghua Wang, Xiaofan Zhang, Deming Chen
The Asia and South Pacific Design Automation Conference (ASP-DAC'24)

[8] **[TECHCON'23] ScaleFlow: High-Level Synthesis for Large Dataflow Applications**
Hanchen Ye, Deming Chen
The Semiconductor Research Corporation (SRC) TECHCON (TECHCON'23)

[9] **[ISPD'23] High-level Synthesis for Domain Specific Computing (Invited)**
Hanchen Ye, Hyegang Jun, Jin Yang, Deming Chen
The International Symposium on Physical Design (ISPD'23)

[10] **[FPGA'23] CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture**
Jinming Zhuang, Jason Lau, **Hanchen Ye**, Zhuoping Yang, Yubo Du, Jack Lo, Kristof Denolf, Stephen Neuendorffer, Alex Jones, Jingtong Hu, Deming Chen, Jason Cong, Peipei Zhou
The ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA'23)

[11] **[TRETS] AutoScaleDSE: A Scalable Design Space Exploration Engine for High-Level Synthesis (Journal)**
Hyegang Jun, **Hanchen Ye**, Hyunmin Jeong, Deming Chen
The ACM Transactions on Reconfigurable Technology and Systems (TRETS)

[12] **[DAC'22] ScaleHLS: a Scalable High-Level Synthesis Framework with Multi-level Transformations and Optimizations (Invited)**
Hanchen Ye, Hyegang Jun, Hyunmin Jeong, Stephen Neuendorffer, Deming Chen
The ACM/IEEE Design Automation Conference (DAC'22)

[13] **[HPCA'22] ScaleHLS: A New Scalable High-Level Synthesis Framework on Multi-Level Intermediate Representation**
Hanchen Ye, Cong Hao, Jianyi Cheng, Hyunmin Jeong, Jack Huang, Stephen Neuendorffer, Deming Chen
The IEEE International Symposium on High-Performance Computer Architecture (HPCA'22)

[14] **[MLBench'21] Being-ahead: Benchmarking and Exploring Accelerators for Hardware-Efficient AI Deployment**
Xiaofan Zhang, **Hanchen Ye**, Deming Chen
The MLSys Workshop on Benchmarking Machine Learning Workloads on Emerging Hardware (MLBench'21)

[15] **[LATTE'21] ScaleHLS: Achieving Scalable High-Level Synthesis through MLIR**
Hanchen Ye, Cong Hao, Hyunmin Jeong, Jack Huang, Deming Chen
The ASPLOS Workshop on Languages, Tools, and Techniques for Accelerator Design (LATTE'21)

[16] **[ICSICT'20] IDLA: An Instruction-based Adaptive CNN Accelerator**
Peng Gao, Zhize Huang, **Hanchen Ye**, Gengsheng Chen

The IEEE International Conference on Solid-State & Integrated Circuit Technology (ICSICT'20)

[17] **[ICCAD'20] DNNExplorer: A Framework for Modeling and Exploring a Novel Paradigm of FPGA-based DNN Accelerator**
Xiaofan Zhang*, **Hanchen Ye***, Junsong Wang, Yonghua Lin, Jinjun Xiong, Wen-mei Hwu, Deming Chen
The ACM/IEEE International Conference on Computer-Aided Design (ICCAD'20)

[18] **[DAC'20] HybridDNN: A Framework for High-Performance Hybrid DNN Accelerator Design and Implementation**
Hanchen Ye, Xiaofan Zhang, Zhize Huang, Gengsheng Chen, Deming Chen
The ACM/IEEE Design Automation Conference (DAC'20)

[19] **[ICSICT'18] A Resource-Sharing & Pipelined Design Scheme for Dynamic Deployment of CNNs on FPGAs**
Hanchen Ye, Gengsheng Chen
The IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT'18)

POSTERS

[1] **[HDR'24] StreamTensor: A Compiler from PyTorch to FPGA for AI/ML Applications**
Hanchen Ye, Deming Chen
NSF Harnessing the Data Revolution (HDR) Ecosystem Conference (HDR'24)

[2] **[DAC'23] ScaleHLS: A Scalable High-Level Synthesis Framework**
Hanchen Ye, Deming Chen
Ph.D. Forum of the ACM/IEEE Design Automation Conference (DAC'23)

[3] **[ICCAD'22] vHLS: Verifiable and Efficient High-Level Synthesis**
Hanchen Ye, Deming Chen
Student Research Contest (SRC) of the ACM/IEEE International Conference on Computer-Aided Design (ICCAD'22)

[4] **[A3D3'22] ScaleFlow: Scalable High-Level Synthesis for Large Dataflow Applications**
Hanchen Ye, Deming Chen
Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Annual Meeting (A3D3'22)

[5] **[DAC'22] PolyAIE: A Dataflow Compiler for Heterogeneous Compute Platforms**
Hanchen Ye, Deming Chen
Young Fellow Program of the ACM/IEEE Design Automation Conference (DAC'22)

PATENTS

[1] **[CN] Special-shaped Pipeline Design Method Based on FPGA Local Dynamic Reconstruction Technology**
Gengsheng Chen, **Hanchen Ye**, Siyu Ni, Chao Huang
China Patent CN108228966B (CN)

TALKS AND TUTORIALS

[1] **[FPGA'24] ScaleHLS-HIDA: From PyTorch/C++ to Highly-optimized HLS Accelerators (Tutorial)**
Hanchen Ye, Junhao Pan, Deming Chen
The ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA'24)

[2] **[Intel] HIDA: A Hierarchical Dataflow Compiler for High-Level Synthesis (Invited)**
Intel High-level Design (HLD) Reading Group (Intel)

[3] **[HACC] HIDA: A Hierarchical Dataflow Compiler for High-Level Synthesis (Invited)**
AMD-UIUC Center of Excellence Seminars (HACC)

[4] **[UIUC] Scalable High-Level Synthesis for AI Accelerator Design and Verification**
UIUC Ph.D. Preliminary Exam (UIUC)

[5] **[UIUC] MLIR, ScaleHLS, and HIDA (Guest Lecture)**
UIUC ECE527 (System-On-Chip Design) Guest Lecture (UIUC)

[6] **[Google] ScaleFlow: Scalable High-Level Synthesis for Dataflow Applications (Invited)**
Google X Journal Club (Google)

[7] **[UIUC] MLIR and ScaleHLS (Guest Lecture)**
UIUC ECE527 (System-On-Chip Design) Guest Lecture (UIUC)

[8] **[Intel] Hardware Compilation with MLIR and CIRCT (Invited)**
Intel Strategic CAD Labs (SCL) Tech Presentation (Intel)

[9] **[FPGA'22] ScaleHLS: A New Scalable High-Level Synthesis Framework on Multi-Level Intermediate Representation (Invited)**
The FPGA Workshop on Open-Source Source-to-Source Transformation for High-Level Synthesis (FPGA'22)

[10] **[Gatech] Compilers for Domain-Specific Accelerators (Guest Lecture)**
Gatech ECE6100/CS6290 (Advanced Computer Architecture) Guest Lecture (Gatech)

[11] **[UIUC] ScaleHLS: A New Scalable High-Level Synthesis Framework on Multi-Level Intermediate Representation (Invited)**
UIUC CS Compiler Seminar (UIUC)

[12] **[UIUC] ScaleHLS: A New Scalable High-Level Synthesis Framework on Multi-Level Intermediate Representation (Guest Lecture)**
UIUC ECE527 (System-On-Chip Design) Guest Lecture (UIUC)

[13] **[CIRCT] FSM (Finite-State Machine) Dialect in CIRCT (Invited)**
Circuit IR Compilers and Tools Open Meeting (CIRCT)

[14] **[XACC] ScaleHLS: Scalable High-Level Synthesis through MLIR (Invited)**
Xilinx Adaptive Compute Clusters Tech Talk Series (XACC)

[15] **[CCF] CIRCT: The Next-Generation Open-Source Hardware Compilation Framework based on MLIR (in Chinese) (Invited)**
CCF Agile Hardware Development and Open-Source EDA Forum (CCF)

[16] **[UCSC] ScaleHLS: Scalable High-Level Synthesis through MLIR (Invited)**
UCSC Hardware Systems Collective (HSC) Seminar (UCSC)

[17] **[UIUC] HybridDNN: A Framework for High-Performance Hybrid DNN Accelerator Design and Implementation**
UIUC Ph.D. Qualifying Exam (UIUC)

[18] **[OSDT] Handshake-based High-Level Synthesis in CIRCT (in Chinese) (Invited)**
Open-Source Development Tools Open Meeting (OSDT)

[19] **[CIRCT] Handshake-based High-Level Synthesis in CIRCT (Invited)**
Circuit IR Compilers and Tools Open Meeting (CIRCT)

TEACHING SERVICES

- **Teaching Assistant**
  UIUC ECE527: System-On-Chip Design *Fall 2022, Fall 2023*

- **Guest Lecture**
  UIUC ECE527: System-On-Chip Design *Fall 2021, Fall 2022, Fall 2023*
  Gatech ECE6100/CS6290: Advanced Computer Architecture *Fall 2021*

PROFESSIONAL SERVICES

- **Program Committee**
  ASPLOS Workshop on Languages, Tools, and Techniques for Accelerator Design (LATTE) *2022, 2023*
- **Reviewer**
  Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD) *2021 - 2023*
  Springer Neural Processing Letters (NEPL) *2021*
- **External Reviewer**
  International Conference on Computer-Aided Design (ICCAD) *2023*
  International Symposium on Field-Programmable Custom Computing Machines (FCCM) *2022 - 2024*
  International Symposium on Field-Programmable Gate Arrays (FPGA) *2021, 2023, 2024*