Mengtian Yang

mengtian.yang@utexas.edu | Google Scholar | Github | LinkedIn

Research Interests

- **System/Architecture/Circuit Co-design**
- **AI/COP Accelerators**
- **MLSys/AI Infrastructure**

Education

The University of Texas at Austin *2021 – 2026*

- Ph.D. and Master's degree in Electronic and Computer Engineering, advised by Professor Jaydeep Kulkarni
- **Focus on system/architecture/circuits co-design for AI and combinational optimization problems**

Shanghai Jiao Tong University (SJTU) *2017 – 2021*

- Bachelor's degree in Computer Science, Zhiyuan Honor Program
- **Class representative of ACM Honor Class**, which is an elite CS program for **top 5%** talented students in SJTU

Work Experience

Research Scientist @ ByteDance Seed *2026-Present*

- **Research Scientist** at AI Infrastructure team

Research Scientist Intern @ ByteDance Seed *2024 Summer / 2025 Summer*

- **Designed the communication backend for a Large Language Model simulator**, enabling simulation of collective communication across multi-hierarchical clusters with diverse topologies and algorithms.
- **Developed a communication overlap slowdown model with link-congestion awareness**, capable of analyzing concurrent transmissions and predicting performance degradation due to physical network congestion.
- **Benchmarked and evaluated the simulator against multiple open-source LLM simulators**, performing comparative studies under various cluster configurations and workloads.

Graduate Intern @ Design and Signoff Group, Cadence *2023 Summer*

- **Built an FPGA-based hardware acceleration solution for Early Global Routing**, leveraging hardware-based parallel prefix algorithms to accelerate maze-routing problems.

Graduate Intern @ E-core Custom Memory Team, Intel *2022 Summer*

- **Built a Python-based timing predictor for custom memory macros**, integrating existing timing data into an SQLite database and enabling predictions of timing characteristics for unseen memory configurations.

Teaching Assistant @ ECE Department, UT Austin *2021 - 2026*

- Teaching Assistant for Operating System / Parallel Computer Architecture / Digital Logic / Very-large-scale Integration (VLSI) / ASIC Design Lab I (chip tapeout course) / ASIC Design Lab II (chip measurement course)

Research Experience

Circuit Research Lab (CRL), The University of Texas at Austin *2021 - 2026*

- **Designed a ternary-aware LLM accelerator featuring lossless weight compression and reconfigurable LUT tensor cores.** By utilizing a symmetric Gray ternary codec and decompress-free computation, the design minimizes memory bottlenecks. Fabricated in 16nm FinFET, it achieves 42.9 TOPS/W/mm^2 and 0.862 TOPS/mm^2, delivering an 11.2x PPA improvement and 3.55x higher area efficiency compared to the state-of-the-art accelerators.
- **Designed the first 3D Gaussian Spaltting accelerator that achieves a fully parallelized rendering pipeline**. With depth speculation, preprocessing, sorting, and rasterization can be parallelized without noticeable image quality degradation in real-scenario 3D Gaussian viewing applications. Achieves up to 2.3x PPA improvement and 2.9x energy savings compared with the state-of-the-art accelerator GSCore.
- **Designed the first Compute-in-memory solver chip for Integer Linear Programming**. It supports arbitrary precision COP variables/coefficients and features all-digital in-memory computation, achieving 234x cycle and 10^5x energy saving compared with commercial solvers.
- **Designed a 3D eDRAM TensorCore Architecture which brings computation logic to 3D** based on BEOL and monolithic 3D technology. Proposed out-product Tensorcore-style dataflow to enable larger BL swing for accumulation, which achieves 119 GOP/mm^2 computation density.

HAN Lab, Massachusetts Institute of Technology *2020 - 2021*

- **Build a hardware-software co-design searching system**. It searches hardware architectures(HAS), dataflow mappings and neural networks(NAS) simultaneously to improve accuracy, decrease latency and energy. Achieve 2.5x speedup, 1.8x energy saving and 2.7% ImageNet accuracy improvement compared with running ResNet-50 on Eyeriss.

Emerging Parallel Computing Center, SJTU *2019-2020*

- **Develop a Google TPUv2 simulator with a complete ISA for SIMD-systolic array fused system** and a novel memory/dataflow for efficient convolution, reducing irregular memory accesses by 90% versus the im2col algorithm.

Publications

- **Charon: A Unified and Fine-Grained Simulator for Large-Scale LLM Training and Inference** (First Author) *MLSys 2026*
- **TACC: A 42.9 TOPS/W/mm^2 Ternary/INT/FP LLM Accelerator with Lossless Weight Compression** (First Author) *CICC 2026*
- **GSAcc: Accelerate 3D Gaussian Splatting via Depth Speculation and Gaussian-centric Rasterization** (First Author) *DAC 2025*
- **Data Movement-Aware, Ping-Pong Ising Machine Supporting Full Connectivity and Variable Bitwidths** (Co-author) *ESSERC 2024*
- **CILP: An Arbitrary-bit Precision All-digital Compute-in-memory Solver for Integer Linear Programming Problems** (First Author) (Invited paper) *ASSCC 2024*
- **CILP: An Arbitrary-bit Precision All-digital Compute-in-memory Solver for Integer Linear Programming Problems** (First Author) *CICC 2024*
- **Vecim: A 289.13GOPS/W RISC-V Vector Co-Processor with Compute-in-Memory Vector Register File for Efficient High-Performance Computing** (Second Author) *ISSCC 2024*
- **Power and EM Side Channel Attack resilient AES-128 Core with Round-Aligned Globally-Synchronous-Locally-Asynchronous Operation based on Tunable Replica Circuits** (Co-author) *ISSCC 2024*
- **Snap-SAT: A One-Shot, Energy-Performance-Aware, All-Digital, Compute-in-Memory Solver for Large-Scale Hard Boolean Satisfiability Problems** (Second Author) *ISSCC 2023*
- **A 118 GOPS/mm^2 3D eDRAM TensorCore Architecture for Large-scale Matrix Multiplication** (Co-First Author) *HiPC 2023*
- **CIMGN: An Energy-efficient All-digital Compute-in-memory Graph Neural Network Processor** (Second Author) *ESSCIRC 2023*
- **A GNN Computing-in-Memory Macro and Accelerator with Analog-Digital Hybrid Transformation and CAM enabled Search-reduce.** (Co-author) *CICC 2023*
- **Ising-CIM: A Reconfigurable and Scalable Compute within Memory Analog Ising Accelerator for Solving Combinatorial Optimization Problems** (Co-author) *JSSC 2022*
- **NAAS: Neural Accelerator Architecture Search** (Co-First Author) *DAC 2021*
- **Characterizing and Demystifying the Implicit Convolution Algorithm on Commercial Matrix-Multiplication Accelerators** (Second Author) *IISWC 2021*

Skills

PDKs	Tapeout (Frontend + Backend) Experience: TSMC 65nm / 40nm / 16nm FinFET, Intel16 FinFET
EDA Tools	Candence Virtuoso/Innovus/Genus, Synopsys DC/ICC/PT/FM/ICV/VCS, Xilinx Vivado, Siemens Calibre
Programming	Python / C / C++ / CUDA / Verilog / Tcl / Bash / Latex / C# / Java / Matlab / HLS / Scala / Expect / SQL / JS
Frameworks	PyTorch / TensorFlow / Megatron / vLLM / Nsight / NCCL / HuggingFace / Docker / Chisel / Unity

Honors And Awards

- **Engineering Fellowship** in The University of Texas at Austin *2021 - 2022*
- **Outstanding Graduates in Shanghai** *2021*
- **Zhiyuan Outstanding Scholarship**, **top 0.07%** of 17,000 students in SJTU *2021*
- **Sunshine Leader Scholarship**, **top 1.5%** of 17,000 students in SJTU *2018*
- **Meritorious Winner** of **Mathematical Contest in Modeling** , **top 7%** of 10,670 international teams *2018*
- **Zhiyuan Honorary Scholarship**, **top 5%** of 17,000 students in SJTU *2017-2020*