YONGHAO TAN

Ph.D. Candidate in Electronic and Computer Engineering, The Hong Kong University of Science and Technology (HKUST)
HKUST, Clear Water Bay, Hong Kong, China | +852 53946864 | ytanaz@connect.ust.hk
yonghaot1017@gmail.com | yonghao-tan.github.io | ORCID 0000-0001-5372-5863

EDUCATION

Ph.D. in Electronic and Computer Engineering *Sept. 2023 - Present*

The Hong Kong University of Science and Technology (HKUST) *Hong Kong, China*

Supervisor: Prof. Tim Kwang-Ting Cheng

B.E. in Microelectronics *Sept. 2019 - Jun. 2023*

Southern University of Science and Technology (SUSTech) *Shenzhen, Guangdong, China*

Supervisor: Prof. Fengwei An

Overall GPA: 3.77 / 4.0 | Weighted Average: 90.38 | Rank: 11 / 77

RESEARCH EXPERIENCE

5nm UCIe-Enabled Multi-Chiplet Generalizable Rendering Processor *Mar. 2024 - Sept. 2025*

AI Chip Center for Emerging Smart Systems (ACCESS), Hong Kong, China

- For a 5nm four-chiplet generalizable neural rendering (GeNeRF) processor, designed a Universal Chiplet Interconnect Express (UCIe) cross-die cache and source-view placement flow to keep reused features on package.
- Built a coarse-to-fine sparse rendering flow that skips low-value rays, prunes low-impact source views, and shares fine-stage work across neighboring tiles to reduce computation.
- Added a patch-level super-resolution schedule that routes simple regions to lightweight upsampling and keeps full rendering on detail-sensitive regions.
- Measured 91.43 TOPS/W, 55.43 FPS, 0.29 uJ/pixel, and 95.78% lower external-memory access in silicon.

55nm ReRAM-on-Logic Stacked LLM Accelerator *Apr. 2024 - Aug. 2025*

AI Chip Center for Emerging Smart Systems (ACCESS), Hong Kong, China

- For a 55nm edge large language model (LLM) accelerator, developed a two-stage local rotation flow that stabilized 4-bit weight and 8-bit activation quantization for the target model.
- Built a stacked resistive RAM (ReRAM) memory path with blockwise codebooks so draft-model weights could be reconstructed on package instead of being repeatedly fetched from external memory.
- Implemented adaptive parallel speculative decoding and an out-of-order scheduler that changed drafting strategy from verification feedback and overlapped compute, memory, and communication.
- Measured 14.08 to 135.69 token/s and 4.46x to 7.17x speedup over vanilla speculative decoding, using 8MB in-stack storage over 2048 face-to-face bumps at 25.6 GB/s.

28nm CNN-Transformer Accelerator for Semantic Segmentation *Nov. 2021 - Sept. 2024*

AI Chip Center for Emerging Smart Systems (ACCESS), Hong Kong, China

- For a 28nm ConvFormer and SegFormer accelerator, built a hybrid attention engine that used linear attention for most tiles and kept a small number of full-attention tiles for accuracy.
- Designed a layer-fusion schedule with key/value and weight reuse so fused attention and convolution blocks could share buffered data instead of reloading it from external memory.
- Implemented a cascaded feature-map pruning flow for the segmentation head, using expansion, mask-based pruning, and density restoration to remove low-value compute.
- Measured 0.22 uJ/token on SegFormer-B0 and up to 52.90 TOPS/W in 28nm silicon, with 91.10% less segmentation-head computation.

PUBLICATIONS

A 5nm 91.43 TOPS/W 4-Chiplet Generalizable-Rendering-Processor with UCIe-Enabled Cross-Die-Cache and Balance-Aware Progressive Multi-Level Sparsity
Tan, Y.*, Ma, S.*, Dong, P., Luo, P., Lei, Z., Lu, W., Ying, G., ... & Cheng, K. T.
2026 IEEE Custom Integrated Circuits Conference (CICC), IEEE.

A 14.08-to-135.69Token/s ReRAM-on-Logic Stacked Outlier-Free Large-Language-Model Accelerator with Block-Clustered Weight-Compression and Adaptive Parallel-Speculative-Decoding
Dong, P., Tan, Y., Liu, X., Luo, P., Liu, Y., Pang, D., Ma, S., ... & Cheng, K. T.
2026 IEEE International Solid-State Circuits Conference (ISSCC), IEEE.

A 28nm 0.22uJ/Token Memory-Compute-Intensity-Aware CNN-Transformer Accelerator with Hybrid-Attention-Based Layer-Fusion and Cascaded Pruning for Semantic-Segmentation
Dong, P.*, Tan, Y.*, Liu, X., Luo, P., Liu, Y., Liang, L., ... & Cheng, K. T.
2025 IEEE International Solid-State Circuits Conference (ISSCC), IEEE.

APSQ: Additive Partial Sum Quantization with Algorithm-Hardware Co-Design
Tan, Y.*, Dong, P.*, Wu, Y., Liu, Y., Liu, X., Luo, P., Liu, S. Y., Huang, X., Zhang, D., Liang, L., & Cheng, K. T.
2025 62nd ACM/IEEE Design Automation Conference (DAC), IEEE.

Genetic Quantization-Aware Approximation for Non-Linear Operations in Transformers
Dong, P.*, Tan, Y.*, Zhang, D., Ni, T., Liu, X., Liu, Y., ... & Cheng, K. T.
2024 61st ACM/IEEE Design Automation Conference (DAC), IEEE.

A Reconfigurable Coprocessor for Simultaneous Localization and Mapping Algorithms in FPGA
Tan, Y.*, Deng, H.*, Sun, M., Zhou, M., Chen, Y., Chen, L., ... & An, F.
IEEE Transactions on Circuits and Systems II: Express Briefs, 70(1), 286-290, 2022.

A Reconfigurable Visual-Inertial Odometry Accelerator with High Area and Energy Efficiency for Autonomous Mobile Robots
Tan, Y.*, Sun, M.*, Deng, H., Wu, H., Zhou, M., Chen, Y., ... & An, F.
Sensors, 22(19), 7669, 2022.

** Authors marked with an asterisk contributed equally to the corresponding work.*

HONORS AND AWARDS

Best Teaching Assistant Award, Department of Electronic and Computer Engineering, HKUST *Aug. 2025*

Outstanding Graduate (School Level), SUSTech *May 2023*

First-Class Outstanding Students Scholarship with the highest score *Sept. 2022*

Undergraduate Innovation and Entrepreneurship Training Program *Apr. 2022*

Shenzhen Longsys Electronics Company Award (Top 2% in the School of Microelectronics) *Dec. 2021*

First Prize, 2021 National College Students FPGA Innovation Design Competition (Top 22 out of 1,341 teams) *Dec. 2021*

First Prize, 2021 International Competition of Autonomous Running Robots (1st place out of 34 finalist teams) *Oct. 2021*

FUNDING AND SUPPORT

Postgraduate Studentship (PGS) Award in HKUST *Sept. 2023 - Present*

Undergraduate Innovation and Entrepreneurship Training Programs (Provincial Level) *Apr. 2022*

Guangdong College Students' Scientific and Technological Innovation (Provincial Level) *Jul. 2021*

SKILLS

Research Interests: Software/Hardware Co-Design, Model Compression, 3D Processing

Programming Languages: C, C++, Java, Python, SystemVerilog, Verilog HDL, VHDL

Professional Software: AutoCAD, Cadence, Design Compiler, IC Compiler II, MATLAB, Multisim, Silvaco

Languages: English (fluent), Mandarin (native), Cantonese (native)