Pingcheng Dong (董平成)

HKUST, Hong Kong, China
pdongaa@connect.ust.hk
+852 6234 7165



Education

2022-present **PhD in Electronic and Computer Engineering**, *The Hong Kong University of Science and Technology (HKUST), Hong Kong, China.*, supervised by Prof. Tim Kwang-Ting CHENG

2018-2022 **B.E. in Microelectronics Science and Engineering**, *Experimental Class, School of Microelectronics, Southern University of Science and Technology (SUSTech), Shenzhen, China*, GPA: 3.84/4.0, Rank: 4/93

Research Interests

AI Accelerators, Model Compression, Software-Hardware Co-Design, and AI Agents for AI Chip Design.

Research Experience

2026-2027 **A 28nm 3D-LPU with DRAM Hybrid Bonding for Long-Context LLM/VLM Reasoning**

- Designed a general token pruner and merger for KV-cache and visual token pruning, with run-time DRAM bank management and sparsity-aware static compilation.
- Proposed a prefill-decode disaggregated architecture with sparse DRAM-bank connections to reduce logic area.
- Utilized on-chip thermal sensors to develop thermal-aware dynamic DRAM bandwidth management with thermal-sparsity-bandwidth co-optimization via reinforcement learning. *(Tape-out scheduled for Sept. 2026.)*

2025-2026 **A 22nm VLM Accelerator with Workload-Aware DVFS for AI Agents**

- Designed a coarse-grained-header dual-rail power switch and an on-chip DVFS controller, overlapping the transition to cut voltage/frequency switching time to $< 5\,\mu$s.
- Proposed an approximate token pruner with sparsity-aware subgraph optimization in the compiler.
- Leveraged reinforcement learning to let the agent self-tune the voltage/frequency table and chain-of-thought length at runtime, minimizing end-to-end latency and power under an accuracy constraint. *(Taped out.)*

2024-2025 **A 55nm Speculative-Decoding Accelerator with ReRAM-on-Logic 3D Stacking for Edge LLMs**

- Designed a reconfigurable Hadamard Rotation Unit with decomposed FWHT and multi-level data reuse for accurate W4A8 quantization.
- Proposed blockwise vector quantization combining Gumbel-Softmax and QAT, compressing the draft LLM to 2 bits/value and storing all clustered codebooks in stacked ReRAM, with capacity extended across a four-chip

interconnect.
- Collaborated with the compiler team to co-optimize adaptive parallel speculative decoding with a workload-decoupled out-of-order scheduler for high bandwidth and compute utilization. *(Taped out, silicon-verified, and demonstrated.)*

2023-2024 **A 28nm CNN-Transformer Accelerator for Edge Vision Tasks**
- Proposed a hybrid attention mechanism combining vanilla and linear attention; designed the ISA, the SRAM data layout, and the data-reordering module.
- Designed a reconfigurable attention-convolution feature-map pruner based on intermediate feature expansion.
- Co-designed compiler support for KV-weight-reused layer fusion and an on-chip scheduler. *(Taped out, silicon-verified, and demonstrated.)*

2022-2023 **A 28nm Stereo-Matching Processor for Depth Estimation**
- Designed a PE array supporting recursive multi-path cost aggregation in semi-global matching, and led the RTL-to-GDSII physical design.
- Customized an ISP based on a pixel-level pipeline to support image rectification and demosaicing in the raw domain.
- Implemented RGB compression and binary guided image filtering engines for depth refinement. *(Taped out, silicon-verified, and demonstrated.)*

Honors & Awards

2024 HKUST RedBird Academic Excellence Award

2022 Hong Kong PhD Fellowship (HKPFS)

2022 HKUST RedBird PhD Scholarship

2022 Dean's List, SUSTech

2021 China National Scholarship (Top 8 in SUSTech)

Main Publications

* Co-first author

2026 **Pingcheng Dong**, Yonghao Tan, Xuejiao Liu, Peng Luo, Yu Liu, Di Pang, Fengshi Tian, Songchen Ma, Xijie Huang, Shih-Yang Liu, Dong Zhang, Zhichao Lu, Luhong Liang, Chi-Ying Tsui, Fengbin Tu, Liang Zhao, Kwang-Ting Cheng. "RS-LPU: A ReRAM-on-Logic Stacked Large Language Model Accelerator with Software-Hardware Co-Designed Weight Compression and Speculative Decoding", *IEEE Journal of Solid-State Circuits (JSSC), 2026.*

2026 **Pingcheng Dong**, Yonghao Tan, Xuejiao Liu, Peng Luo, Yu Liu, Di Pang, Songchen Ma, Xijie Huang, Shih-Yang Liu, Dong Zhang, Zhichao Lu, Luhong Liang, Chi-Ying Tsui, Fengbin Tu, Liang Zhao, Kwang-Ting Cheng. "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", *IEEE International Solid-State Circuits Conference (ISSCC), 2026.* **(Highlight Paper)**

2025 **Pingcheng Dong***, Yonghao Tan*, Xuejiao Liu, Peng Luo, Yu Liu, Luhong

Liang, Yitong Zhou, Di Pang, Man-To Yung, Dong Zhang, Xijie Huang, Shih-Yang Liu, Yongkun Wu, Fengshi Tian, Chi-Ying Tsui, Fengbin Tu, Kwang-Ting Cheng. "A 28nm 0.22 μJ/Token Memory-Compute-Intensity-Aware CNN-Transformer Accelerator with Hybrid-Attention-Based Layer-Fusion and Cascaded Pruning for Semantic-Segmentation", *IEEE International Solid-State Circuits Conference (ISSCC), 2025.*

2025 Yonghao Tan*, **Pingcheng Dong***, Yongkun Wu, Yu Liu, Xuejiao Liu, Peng Luo, Shih-Yang Liu, Xijie Huang, Dong Zhang, Luhong Liang, Kwang-Ting Cheng. "APSQ: Additive Partial Sum Quantization with Algorithm-Hardware Co-Design", *Proceedings of the 62nd ACM/IEEE Design Automation Conference (DAC), 2025.*

2024 **Pingcheng Dong***, Yonghao Tan*, Dong Zhang, Tianwei Ni, Xuejiao Liu, Yu Liu, Peng Luo, Luhong Liang, Shih-Yang Liu, Xijie Huang, Huaiyu Zhu, Yun Pan, Fengwei An and Kwang-Ting Cheng. "Genetic Quantization-Aware Approximation for Non-Linear Operations in Transformers", *61st ACM/IEEE Design Automation Conference (DAC), 2024.*

2023 **Pingcheng Dong**, Zhuoyu Chen, Ke Li, Lei Chen, Kwang-Ting Cheng, Fengwei An. "A 1920×1080 129fps 4.3pJ/Pixel Stereo-Matching Processor for Pico Aerial Vehicles", *IEEE 49th European Solid State Circuits Conference (ESSCIRC), 2023.*

2022 **Pingcheng Dong***, Zhuoyu Chen*, Zhuoao Li, Ruoheng Yao, Wenyue Zhang, Yangyi Zhang, Lei Chen, Chao Wang, Fengwei An. "Configurable image rectification and disparity refinement for stereo vision", *IEEE Transactions on Circuits and Systems II: Express Briefs, 2022.*

2021 **Pingcheng Dong***, Zhuoyu Chen*, Zhuoao Li, Yuzhe Fu, Lei Chen, Fengwei An. "A 4.29 nJ/pixel stereo depth coprocessor with pixel level pipeline and region optimized semi-global matching for IoT application", *IEEE Transactions on Circuits and Systems I: Regular Papers, 2021.*

2021 **Pingcheng Dong**, Zhuoyu Chen, Zhuoao Li, Yuzhe Fu, Lei Chen, Fengwei An. "A 139 fps pixel-level pipelined binocular stereo vision accelerator with region-optimized semi-global matching", *IEEE Asian Solid-State Circuits Conference (A-SSCC), 2021.*

Skills

Languages: Verilog, Python, C, and Shell scripting

Tools: VCS, Verdi, DC, PT, ICC2, and Innovus

Model compression techniques for CNNs, ViTs, LLMs, and VLMs

Experienced in AI agents for model compression, compiler, and RTL design