YA WANG

ywangmu@connect.ust.hk | (+852) 9816 5561

EDUCATION

The Hong Kong University of Science and Technology (HKUST) | Hong Kong **Sep. 2022 – Sep. 2027 (Expected)**
Ph.D. in Electronic and Computer Engineering; with HKPFS

Nanjing University | Nanjing, China **Sep. 2018 – Jun. 2022**
B.S. in Electronic Information Science and Technology; Minor in Computer Science | GPA: 4.54/5.0; Rank: 3/201

INDUSTRY EXPERIENCE

Huawei Hong Kong Research Center - NPU Systems Group | Hong Kong **Dec. 2025 – Present**
Research Intern

- Contribute to the development of PTO-ISA, Ascend's tile-oriented virtual ISA and programming abstraction, and use it to develop and optimize Ascend 950-series NPU kernels, including FlashAttention-4 and sparse attention.
- Design and maintain an event-driven, trace-based offline profiling platform for evaluating NPU power-management policies, estimating power and performance loss from cycle-accurate or emulation traces.

Alibaba DAMO Academy - Computational Technology Lab | Beijing, China **Jun. 2022 – Mar. 2024**
Research Intern

- Developed an analytical branch-predictor performance model for RISC-V CPUs and correlated it with cycle-accurate simulation (DATE 2024); subsequently explored CPU microarchitecture DSE using explainable fuzzy neural networks and multi-fidelity reinforcement learning (DAC 2024).

RESEARCH EXPERIENCE

GPU/NPU Modeling and AI Accelerator Architecture

TileSim: Unified Tile-Level Performance Modeling for Cross-Architecture AI Accelerators *[Manuscript]*
Accelerator datapaths are diverging across vendors (Blackwell's tcgen05/TMEM path versus Ascend's fused Cube-Vector execution), yet per-architecture simulators cannot answer cross-platform questions: whether a bottleneck is intrinsic to the kernel or an artifact of one datapath. Designed a retargetable framework that expresses kernels as a platform-agnostic tile/dataflow graph and lowers them through per-target dialects into pipeline-granularity traces, with a semantic sidecar verifying structural conformance independently of timing; one abstraction covers both B200 and Ascend 950, validated on FlashAttention-4 against measured B200 latency (7.6% MAPE).

Sim-FA: A GPGPU Simulator for Fine-Grained Asynchronous Pipelines *[TCAD Under Review]*
Identified that Hopper's warp-specialized asynchrony (TMA, WGMMA, mbarrier) breaks the SIMT abstraction underlying existing GPU simulators. Co-designed a WarpGroup-level event model that attributes stalls to producer–consumer overlap, backpressure, and synchronization, reproducing H800 FlashAttention-3 latency within 5.7% MAPE; ablations quantify the performance role of undocumented hardware (L2 request coalescing, sliced-cache hashing), and the companion analytical model exposes why ideal-cache frameworks severely underestimate long-sequence DRAM traffic.

DFAC: Dataflow-Aided Intra-Cluster Coalescer for AI Accelerators *[ASP-DAC Submission]*
Observed that intra-cluster reuse counts are fixed at programming time yet invisible to hardware coalescers, which release merge entries before slower requesters arrive. Co-developed a coalescer that consumes software-registered reuse metadata to hold requests until all expected requesters are collected, delivering up to 2.08× over prior coalescers in bandwidth-constrained FlashAttention and quantifying a DRAM bandwidth-utilization effect absent from prior coalescer studies.

From TensorMap to PageMap: Cache-Aware Semantic Prefetching for Paged LLM Inference *[Ongoing]*
PagedAttention hides logical KV-page reuse behind dynamic physical mappings, limiting conventional prefetchers. Designing a lightweight PageMap ISA hint and hardware path that restores this semantic visibility and converts logical page streams into timely physical prefetch requests without changing the serving abstraction.

CPU Architecture Design, Modeling and DSE

Branch Predictor Performance Analysis Framework *[DATE 2024]*
Recognized that branch predictability is governed by trace regularity that existing entropy metrics capture only partially, and proposed transition entropy, which correlates with miss rate (Pearson 0.91) without any fitting. Established the first analytical model set covering multi-table TAGE internal parameters, enabling branch-predictor DSE that reaches near-Pareto configurations at negligible evaluation cost versus simulation.

RESEARCH EXPERIENCE (CONTINUED)

CPU Architecture Design, Modeling and DSE (Continued)

Explainable Fuzzy Neural Network with Multi-Fidelity RL for Microarchitecture DSE *[DAC 2024]*
Addressed the opacity of black-box DSE algorithms, which prevents designers from auditing or steering the search. Extended the framework with fuzzy neural networks that distill exploration experience into human-readable design rules, trained via multi-fidelity RL that explores on inexpensive analytical models and verifies on RTL simulation, surpassing state-of-the-art baselines under equal simulation budgets.

TAGE-SC-LLM: Large Language Model Hinted Branch Predictor *[HPCA Submission]*
Built on the insight that static code semantics are a robust proxy for dynamic branch correlation, removing the need for per-workload profiling that limits prior sparse predictors. Co-developed BP-LLM, a fine-tuned model that generates per-branch sparse Lasso states offline for runtime use by a hybrid TAGE-SC-L design; improves MPKI by 7.03% and IPC by 3.13% on SPEC CPU2017 Integer at the same 72KB storage budget.

AutoPerf: Automated Construction of RTL-Aligned Cycle-Level CPU Performance Models *[Ongoing]*
Cycle-level CPU models remain hand-built because timing behavior is scattered across RTL control logic, and end-to-end calibration can shrink aggregate error but cannot localize which microarchitectural mechanism is missing. Developing an evidence-grounded workflow that extracts provenance-tracked timing rules from validated RTL, lowers them into gem5 model components, and iteratively validates cycle counts against RTL execution to attribute residual error to missing mechanisms.

Processor and SoC Verification

HiFuzz: Hierarchical RL for Semantic-Aware and Adaptive CPU Fuzzing *[ITC '26, Accepted]*
Identified two structural failures in RL-based fuzzing: constructive generation creates a control-validity dilemma, while aggregate coverage masks starvation of hard-to-reach modules. First-authored a hierarchical fuzzer with dense intrinsic feedback and adaptive module-level rewards; improves Control Register Coverage by 48.9% over Cascade and detects 30 of 60 Encarsia-injected bugs versus 26 for the best baseline.

CURE-Fuzz: Curiosity-Driven Reinforcement Learning for Enhanced Hardware Fuzzing *[ICCAD '25]*
Observed that converged RL policies regenerate near-duplicate tests after coverage feedback turns sparse, capping exploration despite resets. Co-first-authored a hierarchical framework with RND curiosity and multi-metric feedback, raising BOOM MUX coverage from 66% to 75% and total ISA coverage to 96% while surpassing prior fuzzers in bug detection.

SELECTED PUBLICATIONS AND MANUSCRIPTS

- Y. Wang, H. Fan, Z. Liu, X. Zhou, Y. Lyu, J. Xu, and W. Zhang. "HiFuzz: Hierarchical Reinforcement Learning for Semantic-Aware and Adaptive CPU Fuzzing." (IEEE ITC '2026), Accepted.
- H. Fan*, Y. Wang*, X. Zhou, S. Li, B. Zhao, Y. Lyu, J. Xu, and W. Zhang. "CURE-Fuzz: Curiosity-Driven Reinforcement Learning for Enhanced Hardware Fuzzing in Agile Testing." (ICCAD '2025), Invited Paper. (*Equal contribution.)
- C. Lai, Y. Wang, Z. Zhou, W. Zhang, R. Chang, and H. Yao. "DFAC: Dataflow-Aided Intra-cluster Coalescer for AI Accelerators." (ASP-DAC), Submitted.
- X. Sun, Y. Wang, M. Li, C. Bai, W. Li, C. Xu, Z. Xie, W. Zhang, and Y. Xie. "TAGE-SC-LLM: Large Language Model Hinted Branch Predictor." (HPCA), Submitted.
- Y. Wang, H. Fan, S. Li, T. Liang, and W. Zhang. "A Modular Branch Predictor Performance Analysis Framework for Fast Design Space Exploration." (DATE '2024), Published.
- H. Fan, Y. Wang, S. Li, T. Liang, and W. Zhang. "Explainable Fuzzy Neural Network with Multi-Fidelity Reinforcement Learning for Micro-Architecture Design Space Exploration." (DAC '2024), Published.
- H. Song, L. Xu, Y. Wang, M. Wang, and Z. Wang. "HSViT: A Hardware and Software Collaborative Design for Vision Transformer via Multi-level Compression." (ISCAS '2024), Published.
- H. Song*, Y. Wang*, M. Wang, and Z. Wang. "UCViT: Hardware-Friendly Vision Transformer via Unified Compression." (ISCAS '2022), Published. (*Equal contribution.)

AWARDS

Hong Kong PhD Fellowship Scheme (HKPFS); RedBird Scholarship (2022–2024); National Scholarship (3/204, 2020).

TECHNICAL SKILLS

Languages: C, C++, Python, Chisel
AI Acceleration: PyTorch, PTO-ISA, GPU/NPU performance modeling
Toolchains: gem5, Chipyard, Cocotb, Ramulator