Yubin Qin

qinyb98@gmail.com | (+86)18260207071 | Google Scholar Homepage

Education

Tsinghua University, Ph.D in Electronic Science and Technology — Sept 2020 – Present

- Advised by Prof. Shouyi Yin.
- Research field: Energy-efficient and high performance AI processors and architecture.
- Publications: 8 first-author papers in top-tier venues (e.g., ISSCC, ISCA); 20+ papers in total; H-index: 15.
- Project Experience: Participated in 8 national-level research projects and 10+ 22/28 nm tape-outs; designed high-performance neural network computing IP adopted in two commercial chips, with cumulative shipments exceeding 10 million units.
- Awards: ISSCC 2025 Highlight; "Top 10 Advances in China's Chip Science"; nominee for "Top 10 Research Advances in China's Semiconductor Industry".
- Scholarships: National Scholarship; National IC Talent Scholarship; two-time nominee for the Tsinghua Special-Class Scholarship; Deng Jiaxian Scholarship (20 out of 46,000+ students); Future Scholar Award; several research excellence scholarships.

Southeast University, B.Eng. in Electronic Science and Technology — Sept 2016 – Jun 2020

- GPA: 3.9/4.0, Rank 1/161.
- Awards: National Scholarship, President's Scholarship, Baosteel Excellent Student Award (8 out of 40000+ students), Philips Scholarship; recognized as Excellent Graduate and Outstanding Student Leader.

Research Publications (* co–first author)

- **Yubin Qin**, Yang Wang, *et al.*, "AMBER: A Unified Accelerator for Multi-Precision LLM Inference Exploiting Bit-Level Redundancy and Reconfigurability," **DAC 2026**.
  - Proposed a unified lossless compression and GEMM computation architecture for various modern AI precision including FP, BF, MX, and NVFP4.
- **Yubin Qin**, Yang Wang, *et al.*, "An 88.36 TOPS/W Bit-Level-Weight-Compressed LLM Accelerator with Cluster-Aligned INT-FP-GEMM and Bi-Dimensional Workflow Reformulation," **ISSCC 2025**.
  - Designed an LLM accelerator co-optimizing memory compression, mixed-precision computation, and hardware utilization bottleneck, achieving 88.36 TOPS/W on LLaMA-2-7B and offering a new system-level direction for general-purpose LLM accelerators.
  - Recognized as **ISSCC Highlight** in AI Accelerator Session.
- **Yubin Qin**, Yang Wang, *et al.*, "MECLA: Memory–Compute-Efficient LLM Accelerator with Scaling Sub-matrix Partition," **ISCA 2024**.
  - Proposed a scalable sub-matrix partitioning and PSum-reuse–based matrix reorganization scheme for large model inference, reducing memory accesses by 83.6% and computations by 72.2%, and achieving 7,088 GOPS/W across 20 benchmarks.
  - Awarded with **ISCA 2024 Student Travel Grant**
- **Yubin Qin***, Yang Wang*, *et al.*, "A 52.01 TFLOPS/W Diffusion Model Processor with Inter-Time-Step Convolution Attention Redundancy Elimination," **VLSI 2024**.
  - Designed a high-efficiency diffusion model processor that exploits cross-timestep redundancy and semantic-segmentation-based sparse convolution, combined with a bipolar floating-point multiplier, achieving 52.01 TFLOPS/W in 22 nm.
  - Awarded with **VLSI 2024 Student Travel Grant**
- **Yubin Qin**, Yang Wang, *et al.*, "Ayaka: A Versatile Transformer Accelerator with Low-Rank Estimation and Heterogeneous Dataflow," **IEEE JSSC**.
  - Proposed a 28 nm high-efficiency Transformer processor for dynamic sparse attention, which speculatively identifies important tokens, adapts its dataflow to the resulting sparsity pattern, and suppresses ineffectual

bit-level operations in softmax, achieving end-to-end sparse acceleration and 49.7 TOPS/W.

- **Yubin Qin***, Yang Wang*, *et al.*, "FACT: FFN–Attention Co-optimized Transformer Architecture with Eager Correlation Prediction," **ISCA 2023**.
  - Proposed an Eager Correlation Prediction (EP) algorithm–hardware co-design that jointly accelerates QKV, attention, and FFN in Transformers, achieving 3.59× end-to-end layer throughput and 4,388 GOPS/W across 22 benchmarks;
  - This ISCA 2023 paper has been highly cited (100+ citations).
- **Yubin Qin***, Yang Wang*, et al., "A 28nm 49.7TOPS/W Sparse Transformer Processor with Random-Projection-Based Speculation, Multi-Stationary Dataflow, and Redundant Partial Product Elimination," **A-SSCC 2023**
  - Proposed a 28 nm high-efficiency Transformer processor for dynamic sparse attention, which optimizes the sparsity speculation and computation mechanism via random-projection algorithm, achieving an energy efficiency of 49.7 TOPS/W.
- Yushu Zhao*, **Yubin Qin***, et al., "MoBiLE: Efficient Mixture-of-Experts Inference on Consumer GPU with Mixture of Big Little Experts," **ASP-DAC 2026**
  - Developed MoBiLE, an MoE inference framework that mitigates CPU–GPU bandwidth bottlenecks by dynamically routing tokens to large or small experts based on importance and using expert-switch fallback and prefetching to balance GPU HBM and CPU DRAM usage.

Ongoing Projects (as Project Lead)

- **Heterogeneous In-Memory–Digital LLM Accelerator Chip:**
  Developing a heterogeneous in-memory–digital LLM accelerator chip that integrates in-memory computing with digital compute on-chip, enabling parallel KV-cache access and matrix multiplication to substantially improve LLM inference performance.
- **Generalized Quantized LLM Inference Architecture (Arbitrary Fixed-/Floating-Point Precision):**
  Designing a generalized quantized LLM inference architecture with unified, configurable arithmetic and storage formats, supporting mixed-precision computation from INT2 to FP16 and adaptive compression, enabling highly flexible, low-power inference with strong accuracy retention.
- **LLM Inference Architecture with Exponent Redistribution and Shift Sharing:**
  Exploring an LLM inference architecture that leverages floating-point exponent redistribution and shift sharing to reduce redundant multiplications and data movement, achieving jointly optimized compute density and memory-access energy.
- **Ultra-low-latency LLM Inference Architecture with PIM:**
  Developing an ultra-low-latency LLM inference architecture with full-stack optimization, including algorithm, framework, network, micro-architecture and DRAM-based PIM circuits. Achieving lower TBT compared to modern LLM inference chips like Groq LPU and Taalas.

Teaching Experience

Teaching Assistant, Computer Organization and Architecture 2024, 2025 Spring

- Assisted in teaching an undergraduate course on computer architecture and principles, covering instruction, pipelines, and memory hierarchy.
- Designed and graded homework assignments and exams, and provided guidance on course projects.

Academic Service and Engagement

- **Academic Service:**
  Reviewer for top-tier journals in integrated circuits, including *IEEE JSSC*, *TCAS-I*, and *JETCAS*.
  Assisting reviewer for leading conferences in computer architecture and integrated circuits, including *ISCA*, *MICRO*, *HPCA*, *DAC*, *A-SSCC*, and *FPGA*.
- **Academic Engagement:**
  Delivered over 10 invited keynote talks about AI accelerator design, including talks *ICFC Chip Forum*, *"Ph.D. Innovators in Chips"* (Beijing Innovation Center for Future Chips), *"Zhishan Chip Talk"* (Southeast University & National EDA Innovation Center), and the *European MiM Forum on Advanced In-Memory Computing*.