秦育彬

18260207071 qinyb98@gmail.com

教育经历

清华大学，集成电路学院，电子科学与技术，博士 导师：尹首一 2020.09 – 至 今

- **研究方向**：高能效AI处理器架构与芯片设计，高性能LLM推理芯片与架构设计
- **论文发表**：第一作者发表ISSCC,ISCA等集成电路和计算机架构领域顶会顶刊8篇，总计20余篇，H-index 15
  获2025 ISSCC AI Accelerator Highlight（远东地区唯一获奖）；VLSI 2024、ISCA 2024 Travel Grant
- **项目经历**：参与8项国家级科研项目，完成10余次22/28nm流片
  设计高性能神经网络计算IP，落地于2款商业芯片，迄今累计出货超过千万颗
- **奖学金**：曾获国家奖学金，国家集成电路紧缺人才专项奖学金，二次获得清华大学特等奖学金提名奖，邓稼先奖学金（全校20人），清华大学未来学者奖学金，集成电路学院科研优秀奖学金等
- **荣誉与获奖**：曾获ISSCC 亮点工作，中国芯片科学十大进展，中国半导体十大研究进展提名等
- **学术任职**：担任集成电路领域顶级期刊JSSC，TCAS-I，JETCAS审稿人；担任计算机结构和集成电路领域顶级会议ISCA，MICRO，HPCA，DAC，A-SSCC，FPGA协助审稿人
- **学术主题Talk**：ICFC华人芯片研讨会，北京市未来芯片技术高精尖创新中心《博士创芯说》，东南大学与EDA国创中心《至善芯语》，欧洲先进存储计算MiM论坛

东南大学，电子科学与工程学院，电子科学与技术，学士 2016.09 – 2020.06

- **课业成绩**：GPA：3.9/4.0, 排名1/161
- **奖学金与荣誉**：曾获国家奖学金，校长奖学金，宝钢优秀学生奖（全校8人），菲利普奖学金。曾获东南大学优秀毕业生，优秀团干部
- **学生工作**：担任电子学院科技协会副主席兼任技术部部长；担任班级文宣委员

工作经历

字节跳动，AI芯片微架构 实习 2026.04 – 至 今

- **工作内容**：调研探索AI计算集群和芯片方案，设计高性能推理芯片架构与电路。

代表性科研成果(* 为共同一作)

- Yubin Qin, Yang Wang, et al., “AMBER: A Unified Accelerator for Multi-Precision LLM Inference Exploiting Bit-Level Redundancy and Reconfigurability” **DAC 2026(体系结构顶会，已接收未发表)**
  - **研究创新**：本研究提出一种面向LLM推理的多样化混合精度计算负载优化的可重构计算架构和统一无损显存压缩机制，在IEEE标准FP、BF浮点、MX浮点和NV浮点上实现显存和计算效率提升。
- Yubin Qin, Yang Wang, et al., “An 88.36TOPS/W Bit-Level-Weight-Compressed Large Language Model Accelerator with Cluster-Aligned INT-FP-GEMM and Bi-Dimensional Workflow Reformulation” **ISSCC 2025(集成电路顶会)**
  - **研究创新**：本研究提出主动位分配格式(ABAF)，簇对齐混合精度处理单元(CAMP-PE)，和双维度MAC-SFU重构(BMSR)，在存储压缩、混合精度计算与非线性函数计算结构三方面协同创新，显著提升了大语言模型(LLM)推理能效与利用率。芯片在运行LLaMA-2-7B时达到峰值88.36 TOPS/W的能效，为通用LLM加速器提供了新的系统级优化方向
  - **亮点**：本研究被ISSCC AI Accelerator 评为 Highlight（亮点工作），也是唯一获奖的中国团队工作
- Yubin Qin*, Yang Wang*, et al., ”FACT: FFN-Attention Co-optimized Transformer Architecture with Eager Correlation Prediction” **ISCA 2023 (CCF-A，计算机结构顶会)**
  - **研究创新**：提出了一套“急切相关性预测（EP）”的算法-硬件协同方案，在QKV生成之前先用基于领先位检测的加法式预测近似注意力相关性，并据此同时稀疏化/跳过QKV生成与注意力计算，再把行级混

合精度分配到FFN；硬件上设计了独热加法树的低功耗预测单元、基于KV差分序的乱序QKV调度，以及对角存储模式以零开销读出INT4/INT8混合精度数据。由此实现全球首个对QKV、注意力、FFN统一加速的Transformer架构，在22个基准上实现3.59×整层吞吐提升与4388 GOPS/W能效。

  - **亮点：本研究是ISCA 2023高被引工作，被引用超过100次（全会议第三），下载量仅次于TPU v4**

- Yubin Qin, Yang Wang, et al., "MECLA: Memory-Compute-Efficient LLM Accelerator with Scaling Sub-matrix Partition", **ISCA 2024(CCF-A，计算机结构顶会)，获得ISCA Student Travel Grant**
  - **研究创新**：本文提出新型可缩放子矩阵划分方法实现了大模型推理中参数访问与计算双重效率提升。该方法实现不访问完整权重矩阵的情况下重建线性层。另提出基于部分和（PSum）复用的矩阵重组与乘法重关联机制，能够在硬件层面动态调整矩阵映射与计算顺序，实现对内积与外积方向的高效复用。该架构在20项基准测试中实现了83.6%的内存访问减少和72.2%的计算量降低，能效达到7088 GOPS/W
- Yubin Qin*, Yang Wang*, et al., "A 52.01 TFLOPS/W Diffusion Model Processor with Inter-Time-Step Convolution-Attention-Redundancy Elimination and Bipolar Floating-Point Multiplication" **VLSI 2024(集成电路顶会)，获得VLSI Student Grant**
  - **研究创新**：本文提出了一种面向扩散模型的高能效处理器，通过跨时间步的冗余消除大幅提升卷积层与注意力层的计算效率。该处理器引入语义分段稀疏卷积引擎以跳过未变化区域的计算，并通过双极性浮点乘法器在保持精度的同时跳过无效尾数乘法，从而显著降低功耗与计算复杂度。整体架构在22nm工艺下实现峰值能效52.01 TFLOPS/W。
- Yubin Qin, Yang Wang, et al., "Ayaka: A Versatile Transformer Accelerator With Low-Rank Estimation and Heterogeneous Dataflow" **JSSC(集成电路顶刊)**
- 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(集成电路顶会)**
  - **研究创新**：以上两篇论文的创新点在于提出了一种面向动态稀疏注意力的高能效Transformer处理器Ayaka，通过随机投影推测（RPAS）、多驻留数据流（HDPE）和冗余部分积消除乘法器（IPRM）协同优化实现全流程稀疏加速。利用随机投影和可切换驻留模式的数据流实现高效的稀疏Transformer计算，并使用IPRM在比特级减少softmax算子的功耗。该设计在28nm工艺下实现49.7 TOPS/W能效。
- Yushu Zhao*, Yubin Qin*, et al., "MoBiLE: Efficient Mixture-of-Experts Inference on Consumer GPU with Mixture of Big Little Experts" **ASP-DAC 2026**
  - **研究创新**：本文提出了一个名为 MoBiLE 的新型 MoE推理框架，旨在突破传统MoE模型在CPU-GPU带宽受限下的推理瓶颈。MoBiLE的核心思想是通过大-小专家混合机制，根据token的重要性动态分配计算资源。为进一步提升内存与数据传输效率，MoBiLE还设计了专家切换回退与预取机制，可在"大专家"和"小专家"间高效切换，实现对GPU HBM和CPU DRAM资源的平衡利用。

在研项目（作为项目负责人或主导者）

- **极低延迟的LLM推理芯片与系统设计：**
  通过融合大规模芯片设计、KV Cache缓存策略、权重自适应量化策略、硬件流水线设计优化等算法-硬件全栈协同设计，实现极低的超大规模LLM（2T及以上参数MoE模型）的解码推理性能。
- **存内计算-数字计算异构的LLM加速芯片：**
  通过在片上融合存内计算与数字计算，实现KV缓存访问与矩阵乘并行加速，显著提升大模型推理的性能。
- **基于浮点指数重分配和移位共享实现计算和存储优化的LLM推理架构：**
  利用指数重分配与移位共享机制减少冗余乘法与数据搬移，兼顾高计算密度与低访存能耗，实现算力与存储的协同优化。