李博宇

(+86)13931326803 | lby1570975210@gmail.com
https://github.com/lllbbbyyy

研究方向

- 面向**多核/多芯粒 AI 加速器**的**编译优化**、**硬件架构设计**和**性能测试评估**方法研究。
- 面向 AI 计算设备的**硬件建模**与**任务调度**方法研究，聚焦于**延迟**和**功耗预测**模型的构建。

教育背景

- **硕博连读** | **中国科学技术大学** | 计算机技术　　合肥、苏州　　2022.09 – 今
  硕士期间**嵌入式方向综合排名第一，获国家奖学金**
- **本科** | **同济大学** | 计算机科学与技术　　上海　　2018.09 – 2022.06

论文发表

- *Magnifier: A Chiplet Feature-Aware Test Case Generation Method for Deep Learning Accelerators.*
  TCAD 2025 (CCF A)，第一作者，EDA 领域顶刊
  - **背景：**当前深度学习加速器缺乏成体系的性能测试方法，现有测试大多关注正确性测试而忽视硬件性能瓶颈的问题，因此提出一种硬件特性感知的测试用例自动化生成方法。
  - **特性提取与生成空间构建：**深入分析了典型的多芯架构加速器在计算、存储和通信模式上的特征，设计了芯粒特性感知的算子任务集；基于相关性分析，提取了常见模型中对性能影响关键的算子任务集；参照神经架构搜索方法，设计了具备拓扑多样性的模型级用例生成空间。
  - **测试评价指标定义：**提出使用跨设备百分位性能标准差来作为衡量用例质量的评价指标，用于有效量化并筛选暴露出软硬件性能缺陷的高价值测试用例。
  - **基于 GAN 的生成加速：**引入生成对抗网络（GAN）来拟合高差异用例的分布情况，将高质量质量测试用例的生成时间从最开始的数小时降低到秒级。
  - **成果应用：**所提方法可以用于算子优化缺陷检测（如张量切分与任务调度异常）及硬件架构设计评估，成果已作为核心组件集成于科技部重点研发项目中。
- *Compass: Co-Exploration of Mapping and Hardware for Heterogeneous Multi-Chiplet Accelerators Targeting LLM Inference Service Workloads.*
  第一作者，TCAD 在投
  - **背景：**大模型（LLM）推理服务具有高度动态性，主要表现为请求类型混合以及序列长度可变，而现有的硬件设计空间探索（DSE）工作大多针对传统 CNN/Transformer 模型。因此提出支持 LLM 动态性的设计框架，支持对延迟、能耗和硬件货币成本的分析评估，以及对映射和硬件架构的搜索优化。
  - **映射编码表示：**设计了一种基于计算执行图的中间表示，以描述动态负载在异构核心上执行方式，可以灵活表示数据、模型、流水线等多种并行方案。
  - **评估器设计：**提出了一种支持层间流水线的数据访问分析算法来对片上网络和 DRAM 访问进行统计，并以此构建了基于分析模型的评估器。可以评估给定模型和硬件架构的延迟、能耗和硬件货币成本。
  - **映射搜索和硬件架构优化：**分别设计遗传算法和贝叶斯优化来对映射表示和硬件架构参数进行优化。
  - **成果应用：**所构建的评估器提供了可替换的单核评估接口，可灵活支持各种多核硬件架构。评估器也可与诸如 Chunked Prefill 等服务调度策略结合。
- *MultiRuler: A Multi-Dimensional Resource Modeling Method for Embedded Intelligent Systems of Autonomous Driving.*
  TVT 2024 (Trans，JCR 一区)，共同第一作者
  - **背景：**自动驾驶场景中的深度学习任务对计算、内存和功耗有严格要求，需要获知任务在设备上的资源消耗情况进行调度。现有方法大多过度依赖硬件细节，因而提出一种通用的多维资源评估方法。

- **核心方法：**提出并设计了基于神经网络的多维资源建模方法。以待评估任务为基础，在“基本块”粒度上生成具备多样性的变体网络，构建贴近真实场景的训练集。在任务角度，将基本块计算量与参数量提取为二维网格，保留了数据依赖。在硬件角度，提取 Batch Size、计算核心数、运行频率等运行时特征。设计了预测网络，将任务特征与硬件特征融合，实现了端到端的多维资源消耗情况预测。
- **成果应用：**在两大主流边缘平台（NVIDIA AGX Xavier 与寒武纪 MLU220）上进行实测，模型对推理延迟、内存占用和功耗的预测误差均在 10% 以下。

- *Arch2End: Two-Stage Unified System-Level Modeling for Heterogeneous Intelligent Devices.*

  TCAD 2024 (CCF A)，学生第二作者，EDA 领域顶刊
  - **背景：**统一的设备性能建模可用于支持任务调度。然而现有的白盒方法受限于闭源的硬件细节，而黑盒方法对多样任务泛化能力受限，因此提出兼顾架构先验与端到端指标的两阶段建模方法。
  - **多维硬件边界探测：**在第一阶段，利用公开信息从计算架构、存储层级和数据通信三个核心维度对异构设备进行抽象，并针对性地设计了架构感知的基准模型集以刻画了不同未知设备的性能边界。
  - **端到端拟合与降维表示：**在第二阶段，构建了可扩展的模拟网络生成器，通过动态采样收集海量端到端推理指标。进一步引入基于线性拟合的特征降维策略，从复杂的“模型-延迟”数据对中提取出统一的低维设备潜在特征。
  - **成果应用：**在涵盖多种现实世界 GPU 与 NPU 的异构平台上实现了较高的预测精度，在 NAS-Bench-201 和真实世界 DNN 上的延迟预测平均相对误差分别低至 1.7% 和 8.2%。此外，提取的统一设备特征也应用于分布式系统中的组间负载均衡调度，降低集群组间执行延迟的方差。

获奖情况

- 苏州工业园区奖学金 | 研究生 2025.03
- **国家奖学金** | 研究生 2023.10
- 国家励志奖学金 | 本科 2020.12

竞赛情况

- 全国大学生物联网设计竞赛**全国总决赛一等奖** 2021.08
- 中国高校计算机大赛-团体程序设计天梯赛**个人二等奖** 2021.05
- **ACM-ICPC** 国际大学生程序设计竞赛**亚洲区域赛铜奖** 2021.04
- CCF 大学生计算机系统与程序设计竞赛**华东赛区铜奖** 2020.10
- CCF-CSP 计算机软件能力认证 **340 分，前 1.5%** 2020.09

工程项目

- 高能效智能工具链开发技术 | **科技部重点研发项目** 2023.03 – 2026.03
  面向国产算力资源的工具链测试平台，负责融合多芯粒特性的测试评价子系统构建。基于 Jenkins 构建了从测试用例生成、测试用例分发执行到测试结果汇总分析的自动化测试平台。
- 自由组配式模块化互联智能锁 | **国家级**大学生创新训练计划项目**优秀结题** 2019.03 – 2021.03
  负责嵌入式端和服务端程序研发，开发了 NFC、红外、烟雾、人脸检测、远程控制等多个智能锁体模块，定义了锁体与服务器间的通信协议，并且基于 ZeroTier 和 Flask 构建了双向通信服务。