# 马天云

1999-07-15 (+86) 18801000859 mty21@mail.ustc.edu.cn
中国科学技术大学 & 中科院计算所 计算机科学与技术 • 博士 中共党员

我是一名计算机系统结构专业博士，研究方向聚焦于大模型推理加速架构，同时探索智能计算系统、处理器自动设计与具身智能等领域。在 TCAD、MICRO、ASPLOS 等 CCF A 类期刊和会议上发表多篇高水平论文，积累了扎实的数理基础与体系结构设计经验。熟练掌握主流深度学习框架（如 PyTorch）、架构仿真工具，以及 C/C++、Python、CUDA 等编程语言，具备优秀的科研能力和工程实践能力，致力于通过高效、创新的体系结构设计推动智能计算的发展。

个人主页：https://matianyun.github.io/

## 教育背景

| | | |
|---|---|---|
| 2022.09<br>至今 | **中国科学技术大学 & 中科院计算所** • 智能处理器研究中心<br>计算机系统结构 • 工学博士 | 导师: 陈华平, 郭崎, 陈云霁<br>学位论文：面向 Transformer 推理加速的关键架构研究 |
| 2020.09<br>2022.06 | **中国科学技术大学** • 微电子学院<br>电子信息 • 硕士 | 导师: 陈云霁 |
| 2016.09<br>2020.06 | **清华大学** • 信息学院自动化系<br>自动化 • 工学学士 | 导师：张涛<br>学位论文：基于 RGB-D 相机的室内多机器人协同 SLAM 及场景重建 |

## 论文发表

已发表 8 篇论文，其中第一作者 3 篇，合作作者 5 篇；另有 4 篇论文正在投稿。
主要研究方向包括：智能计算系统架构设计、具身智能、处理器自动化设计。

- **T. Ma**, Y. Wen, X. Song, P. Jin, D. Huang, H. Han, *et al.*, “Harmonia: A Unified Architecture for Efficient Deep Symbolic Regression.” *TCAD*, 2024（**CCF-A，计算机体系结构顶级期刊**）
- **T. Ma**, J. Guo, D. Huang, X. Song, Q. Wang, H. Han, *et al.*, “DASA: Distribution-Aware Sparse Attention for Accelerating Diffusion Transformer.” *TCAD*, 2025（**CCF-A，计算机体系结构顶级期刊，已接收**）
- **T. Ma**, T. Zhang, S. Li, “Multi-robot collaborative SLAM and scene reconstruction based on RGB-D camera.” *CAC*, 2020 （**中国自动化大会，EI 检索**）
- Z. Yu, S. Liang, **T. Ma**, Y. Cai, Z. Nan, D. Huang, *et al.*, “Cambricon-llm: A chiplet-based hybrid architecture for on-device inference of 70b llm.” *MICRO*, 2024 （**CCF-A，计算机体系结构旗舰会议**）
- C. Li, D. Huang, P. Jin, **T. Ma**, H. Han, S. Cheng, *et al.*, “AGON: Automated Design Framework for Customizing Processors from ISA Documents.” *TCAD*, 2025 （**CCF-A，计算机体系结构顶级期刊**）
- H. Han, X. Zheng, Y. Wen, Y. Hao, E. Feng, L. Liang, J. Mu, X. Li, **T. Ma**, *et al.*, “TensorTEE: Unifying Heterogeneous TEE Granularity for Efficient Secure Collaborative Tensor Computing.” *ASPLOS*, 2024（**CCF-A，计算机体系结构旗舰会议**）
- Y. Zhao, D. Huang, C. Li, P. Jin, M. Song, Y. Xu, **T. Ma**, *et al.*, “Codev: Empowering llms with hdl generation through multi-level summarization.” *TCAD*, 2025 （**CCF-A，计算机体系结构顶级期刊**）
- P. Jin, Z. Fan, Y. Zhao, Z. Du, H. Guo, Z. Nan, Y. Hao, C. Li, **T. Ma**, *et al.*, “SaaP: Rearchitect SoC-as-a-Processor to Orchestrate Hardware Heterogeneity.” *TCAD*, 2025 （**CCF-A，计算机体系结构顶级期刊**）

## 专利

- 一种深度符号回归加速器及深度符号回归方法. 发明人: 胡杏, 马天云, 靳鹏威. 公布号: CN117421703A (第二发明人)
- 基于芯粒和近存计算的边缘端大语言模型推理加速方法. 发明人：胡杏, 于钟凯, 马天云, 梁胜文, 郭崎, 陈天石. 公开号：CN119476487A (第三发明人)
- 一种多机协同构建三维点云地图的方法、装置和存储介质. 发明人: 张涛, 李少朋, 马天云. 公布号: CN111951397B

(已授权, 第三发明人)

## 科研经历

- **Transformer 推理加速架构研究** **博士课题，2022.11–2025.06**

  针对 Transformer 推理中的自注意力、非线性算子和权重相关算子提出创新优化方案:
  - 提出动态稀疏自注意力机制，降低长序列计算复杂度
  - 设计基于线性近似的统一可配置计算阵列，提升非线性算子效率
  - 开发近存协同计算架构，优化权重数据访存开销

  研究成果发表于 TCAD、MICRO 等顶级期刊会议。
- **基于大模型的芯片设计自动化** **2023.03–2025.06**

  致力于利用大语言模型推动芯片设计自动化:
  - 设计微操作中间表示，弥合指令集文档与硬件描述语言语义鸿沟
  - 提出基于 Verilog 代码提示与多级摘要的指令调优框架
  - 开发 RTLSeek 框架，通过多样性导向强化学习优化 RTL 代码生成

  相关成果发表于 TCAD，并投稿 ICLR 会议。
- **高效 VLA 模型推理** **2025.06–至今**

  面向机器人控制中的视觉-语言-动作模型，提出状态自适应动作空间方法:
  - 设计动态一致性掩码数据收集与训练模块, 开发基于动作掩码模型的推理加速框架, 在 LIBERO 仿真环境中显著提升任务成功率与推理速度

  成果已投稿 ICRA 会议。

## 技能和语言

**体系结构** 熟悉 GPU、CPU、NPU 微架构知识、架构设计流程、架构模拟仿真工具等。
**深度学习** 熟练使用 Pytorch 框架, 熟悉主流神经网络模型和算法, 如 Transformer、CNN 等。
**编程语言** 熟悉常用编程语言 (Python, C++, C), CUDA 编程等。
**语言** **英语** — 专业

## 荣誉奖项

- 2023-2024 年中国科大-中关村研究生基础奖学金
- 2017 年清华大学李衍达励学金
- 2018 年东京 IDC 国际机器人设计大赛亚军

## 其他经历和技能

2024.06 – 2024.07 参与 ISCA 2024 Workshop 组织，负责审稿和学者邀请。
- 官方网站：https://ai4facd.github.io/

2022.02 – 2022.08 中科寒武纪科技股份有限公司 (北京) · 验证部实习生
- 负责 MLU370 计算部件功能验证和模拟器开发。

2019.07 – 2019.09 深圳优必选科技有限公司北京研究所 · 感知组实习生
- 负责四足机器人激光雷达 3D-SLAM 算法开发。

2017.09 – 2019.06 清华大学创客空间协会 · 副会长
- 负责面向全校学生的硬件开发工作坊安排和主讲，负责协会日常设备管理。
- 在校期间参与《制造工程实践》课程项目开发和授课，多次参与本科生课程助教工作。

热爱足球，曾担任清华大学自动化系男足队长，带队获得新生杯冠军。