李彦博

南方科技大学 . 电子科学与技术 . yb1.li@siat.ac.cn . (+86) 185-8577-5502

我是李彦博，目前就读于南方科技大学电子科学与技术专业，师从叶可江研究员。目前研究方向为大语言模型推理优化、分布式系统和边缘计算。

教育背景

2022.09 - 至今	南方科技大学 · 电子科学与技术 硕士
2018.09 - 2022.07	哈尔滨工业大学 · 机械工程 本科

研究成果

论文

FLASH: Low-Latency Serverless Model Inference With Multi-Core Parallelism in Edge **ICPADS 2023.11**

Yanbo Li, Yanying Lin, Shijie Peng, Yingfei Tang, Kejiang Ye

- 解决了边缘计算环境中资源受限情况下深度学习模型推理延迟高的问题
- 利用 CPU 多核并行性、动态调整 CPU 核心数量,优化了资源调度算法，从而实现更强大的弹性计算。
- 在不同流量负载下，平均可以将响应延迟降低 33%，最高降低 75%，同时将吞吐量提高 2.94 倍。

QUART: Model Serving System with Resource Fine-Tune in Pipeline Stages **ICDCS 2024.06**

Yanying Lin, **Yanbo Li**, Shijie Peng, Yingfei Tang, Shutian Luo, Haiying Shen, Chengzhong Xu, Kejiang Ye

- 解决了大语言模型推理过程中，由于请求突发的队列堵塞,以及传统扩缩容方法导致资源利用率低下的问题
- 通过流水线拥塞检测与在线扩缩容机制、利用 CV 传播管理资源、CPU 并行减轻通信和控制流开销应对请求波动。
- 平均响应延迟最多降低 87.1%，goodput 最多提高 2.37 倍，在复杂负载下表现更优。

专利

一种基于 **CPU** 多核并行边缘深度模型的推理加速方法及系统 CN202311776975.9 发明专利

模型驱动的云边端互联集成方法、装置、设备及介质 CN202411756916.X 发明专利

一种标准化云边端互联系统 CN202411485258.5 发明专利

一种工业互联网计算干扰避免方法、存储介质和处理器 CN202311744709.8 发明专利

专业技能

掌握	Python, Kubernetes, Prometheus, Granfana, CET6 569 分
熟悉	Transformer, MOE, vLLM
了解	TensorRT-LLM, 并行计算, 网络通信, Deepspeed/Ray 等分布式系统架构

获奖情况

中国科学院深圳先进技术研究院数字所优秀研究生 2023

中国科学院深圳先进技术研究院云计算研究中心十佳研究生 2023

中国科学院深圳先进技术研究院篮球比赛 季军 2024

兴趣爱好

我平时非常喜欢运动，尤其是球类运动，其中最喜欢的有篮球和乒乓球。无论是和朋友一起打比赛，还是自己练习投篮，都让我感受到运动的乐趣和挑战。此外，我喜欢跑步和听音乐，跑步让我在思考问题时更加清晰，有时还能激发新的灵感；听音乐则能缓解压力，提高长时间工作的耐力，从而更好地投入科研工作。