白有辉 博士（2021届）

电话： (+86) 183-5652-1691

Email： byh0912@mail.ustc.edu.cn

主页： https://youhuibai.github.io/

教育背景

2016.9 - 2021.6	中国科学技术大学	计算机科学与技术学院	计算机系统结构（博士） 导师：许胤龙教授、李诚特任研究员
2012.9 - 2016.6	中国科学技术大学	计算机学院与技术学院	计算机科学与技术（本科）

工作经历

2021.8 - 2022.11	华为2012实验室中央软件院	高级工程师A
2022.11 - 至今	华为2012实验室中央研究院理论研究部	高级工程师A

研究方向

分布式AI框架 AI大模型训练/推理加速方法研究，算法和系统co-design

图神经网络系统 针对大规模图基于采样的图神经网络训练过程中数据加载的优化策略研究

专业技能

编程能力： 熟悉 C/C++，Python，了解 Java、Shell等语言，熟悉Linux系统环境下的编程

系统能力：

- 熟悉主流神经网络计算系统PyTorch、MXNet、TensorFlow，大模型训练框架Megatron、DeepSpeed，主流图神经网络库DGL等，有上述系统平台的源码阅读与开发经历。
- 了解国内外一些开源系统，如深度神经网络系统MindSpore、OneFlow，分布式文件系统Hadoop、Ceph，KV存储系统RocksDB，图数据库Neo4j、ArangoDB等。
- 阅读并修改过部分Linux内核代码，如管理RAID的MD模块。
- 自主设计实现梯度压缩使能的分布式深度学习训练框架HiPress（相关代码待开源）；合作设计实现利用静态缓存、 计算和数据加载流水化等技术的图神经网络训练数据加载器PaGraph。

英语水平： 具备英文科技论文读写能力

项目、科研经历

2022.11 – 至今 AI大模型训练/推理加速方案研究

项目描述： AI大模型训练/推理面临着严重的内存墙、计算墙、通信墙问题，本人在底层算子库层面，设计华为昇腾亲和的大颗粒算子融合方案，应用到多个实际业务场景中：科大讯飞百亿大模型，端到端训练速度提升65%；华为云美图模型，端到端推理QPS性能提升4倍，持平V100；在内存管理层面，提出内存交换感知的静态式内存管理方案，相比SOTA方案，动态内存节省30%，总内存节省16%，并设计自适应的重计算策略，利用空闲内存释放计算资源的压力，使得科大讯飞大模型训练速度提升30%；在并行加速层方面，设计3D并行中Tensor并行维度上的通信隐藏方案，讯飞大模型训练速度提升10%；计算、内存、通信三者相互影响，需要一套自动化的方案针对不同场景巧妙权衡。

2021.7 – 2022.11 针对大规模分布式人工智能训练平台的研究

项目描述： 在人工智能领域，随着模型参数量越来越大，扩展到分布式已成为必然趋势。因为参数量巨大，大模型的训练面临着严重的内存墙、效率墙问题。业界往往采用多维度混合并行（scale out）和异构资源（scale up）来使得大模型训练成为可能，但实际上，扩展到分布式之后不得不在内存使用量、通信数据量、计算效率之间权衡，使得该问题异常复杂。本人通过算子优化、框架并行策略优化、通信优化、混合精度计算、算子异构卸载等手段，在256张华为昇腾910A环境下训练鹏城大圣模型，相比PyTorch+V100训练速度提升20%，推荐网络大模型在单机8卡平台上训练速度提升1倍，关键技术贡献到华为自研AI框架MindSpore中。

2018.1 – 2021.6 针对数据并行分布式深度学习系统的研究

项目描述： 深度神经网络的训练昂贵且耗时，单机的算力和存储力无法满足需求，即便是利用GPU加速，因此扩展到分布式是一个趋势。数据并行策略为一个加速训练的重要技术，本项目旨在不影响训练精度的前提下提升基于数据并行的分布式训练系统的吞吐率。

2019.10 – 2021.1　针对分布式图神经网络训练方面的研究

项目描述： 近年来提出的图神经网络，作为图计算领域的典型代表，将图中的结构和属性信息与深度学习中的特征相结合，已被证明在许多与图相关的任务上具有令人信服的表现。本项目旨在不影响训练精度的前提下提升图神经网络训练的吞吐率。

2017.6 - 2017.12　针对事务型数据库写日志及恢复问题的研究

项目描述： 在数据库、文件系统等系统中，错误失效是很常见的，这种错误失效将破坏系统的一致性、可用性，目前大多数系统采用Write-ahead Log(WAL)的方式来避免。但是单一的、共享的日志文件将会成为整个系统性能的瓶颈，尤其是在多核多线程的应用场景下。一个合理的解决方案是将日志文件分布式化，但因为其维护的复杂性以及跨日志的依赖关系，分布式日志的性能往往比单一日志的性能还要差。

获得奖励

工作阶段： 中央研究院2023H1创新先锋一等奖；
2023上半年计算软件平台部优秀个人高质量交付奖；
总裁嘉奖令：2022计算会战930阶段；2022钱江会战530阶段；2023昇腾产品会战630阶段；

博士生阶段： 安徽省优秀毕业生；
ACE SIGCSE 优秀博士论文奖；

研究生阶段： 中国科大-环球数码奖学金，研究生学业奖学金一等奖。

本科生阶段： 连续四年国家励志奖学金；三年优秀学生奖学金铜奖，一年银奖。

科研成果

1. **Youhui Bai,** Cheng Li and Yinlong Xu. "Fast Logging and Recovery Support for Transactional Databases", **USENIX SOSP (CCF推荐A类会议) 2017 Poster**
2. Zhipeng Li**,** Yinlong Xu, Yongkun Li, Chengjin Tian and **Youhui Bai**. "PDS: An I/O-efficient Scaling Scheme for Parity Declustered Data Layout." **IEEE ICPP (CCF推荐B类会议) 2017**
3. **Youhui Bai,** Cheng Li, Zhiqi Lin, Yufei Wu, Youshan Miao, Yunxin Liu and Yinlong Xu. "Efficient Data Loader for Fast Sampling-based GNN Training on Large Graphs", **GNNSys workshop 2021**
4. **Youhui Bai,** Cheng Li, Zhiqi Lin, Yufei Wu, Youshan Miao, Yunxin Liu and Yinlong Xu. "Efficient Data Loader for Fast Sampling-based GNN Training on Large Graphs", **IEEE TPDS (CCF推荐A类期刊) 2021**
5. **Youhui Bai,** Cheng Li, Quan Zhou, Jun Yi, Ping Gong, Feng Yan, Ruichuan Chen and Yinlong Xu."HiPress: A High-Performance, Compression-Aware Framework for Data Parallel DNN Training", **USENIX SOSP (CCF推荐A类会议) 2021**
6. Peng Liang, Yu Tang, Xiaoda Zhang, **Youhui Bai**, Teng Su, linbo qiao, Zhiquan Lai, Dongsheng Li. "A Survey on Auto-Parallelism of Neural Networks Training", **IEEE TPDS (CCF推荐A类期刊) 2023**
7. Quan Zhou, Haiquan Wang, Xiaoyan Yu, Cheng Li, **Youhui Bai**, Feng Yan, Yinlong Xu. "MPress: Democratizing Billion-Scale Model Training on Multi-GPU Servers via Memory-Saving Inter-Operator Parallelism", **IEEE HPCA (CCF推荐A类会议) 2023**
8. Hao Wu, Shiyi Wang, **Youhui Bai (通讯作者)**, Cheng Li, Quan Zhou, Jun Yi, Feng Yan, Ruichuan Chen, Yinlong Xu. "A Generic, High-Performance, Compression-Aware Framework for Data Parallel DNN Training", **IEEE TPDS (CCF推荐A类期刊) 2023**

个人自评

通用能力： 本科至博士阶段持续担任班长，本科期间担任校五星级社团国旗护卫队社长、校爱心思源社社长、院科技部部长，组织过多次班级和社团活动，做事认真负责，具备很强的团队协作能力。曾担任代数结构、图论、编译原理、数据结构、数据库、组合数学等课程助教，具备扎实的计算机基础知识和良好的表达沟通能力。

学习兴趣： 对系统的研究与开发充满热情，认为改进甚至重建系统是一件很酷的事情。曾在做HiPress项目时，跟随导师从零起步，做过机器采购、系统安装、框架搭建、环境配置、系统维护、集群管理、深度学习平台开发等一系列的事情，学习能力强，认真踏实有责任心。

业余爱好： 兴趣爱好广泛，喜欢羽毛球、乒乓球、游泳、跑步等运动项目，组织实验室小伙伴一起参与校园129长跑活动和每周的羽毛球团建活动；爱好听音乐、看电影、读文学书等放松身心的休闲方式，良好的身体条件和愉悦的心情状态是做系统研究的基础和本钱。