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/

个人简介

白有辉，中国科学技术大学计算机科学与技术学院特任副研究员，2021年博士毕业于中国科学技术大学计算机科学与技术专业。主要研究方向为分布式AI大模型训练和推理系统、大模型算法和系统联合创新方向、图神经网络系统等。目前已发表论文10篇，包括计算机系统领域的国际核心会议和期刊ACM SOSP、IEEE TPDS、IEEE HPCA、AAAI等。申请国家发明专利10项、技术秘密1项。相关研究成果荣获ACE SIGCSE优秀博士论文奖、2024世界人工智能大会青年优秀论文奖等。在华为公司工作期间，作为项目经理多次参与公司级攻关项目，解决基于国产昇腾芯片做大模型训推加速的难题，荣获火花奖一等奖、治水攻关特等奖、中央研究院创新先锋一等奖、计算产品线高质量交付奖、总裁嘉奖令等，作为核心骨干获得公司金牌团队奖。

教育&工作经历

2025.6 - 至今	中国科学技术大学	计算机科学与技术学院	特任副研究员
2022.11 - 2025.6	华为技术有限公司	2012实验室中央研究院	主任工程师
2021.8 - 2022.11	华为技术有限公司	2012实验室中央软件院	高级工程师A
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）

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++，CUDA，Ascend C，Python，熟悉Linux系统环境下的编程

系统能力：

- 熟悉大模型训练和推理主流系统，包括：计算系统PyTorch、TensorFlow，业界常用训练框架Megatron、DeepSpeed，推理框架vLLM，国内大模型训练框架MindSpeed，推理框架MindIE、TorchAiR，有上述系统平台的源码阅读与开发经历。
- 深度参与华为公司大模型训练框架MindSpeed、推理框架MindIE的开发和调优，设计并开发过并行策略、内存管理策略等模块。
- 自主设计实现并开源梯度压缩使能的分布式深度学习训练框架HiPress；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

部分项目经历

2023.12 – 至今 AI大模型推理加速方案研究

承担角色： 项目经理，团队10+博士生，从大模型推理软件栈角度分解问题并设计创新方案

问题描述： 大语言模型增量推理分prefill和decoding两个阶段，随着推理序列的变长，Transformer模型核心单元Attention在两个阶段的占比均增加，占整个端到端推理时间的30%以上。Attention在prefill阶段表现出计算密集型的特征，而在decoding阶段则是访存密集型，每次decode一个词，都需要从GPU global memory加载所有kv cache，kv cache的数据搬移开销和存储开销，与序列长度和请求的任务数呈线性递增关系。

主要工作： Attention作为大模型的核心单元，我们从硬件亲和的算子优化、有损压缩算法等角度出发做方案设计，以降低其在prefill和decoding阶段的时间占比：

1. 针对Attention融合算子在昇腾芯片上计算不高效的问题，我们根据昇腾硬件的架构特征，结合Attention的数学计算特点，提出高效的数据分块、指令异步流水、向量类操作转矩阵乘等方法，使得在昇腾910A和910B两个代际的芯片上，Attention算子计算效率均能达到基于竞品硬件的FlashAttention2的性能，有效助力大语言模型、多模态模型的业务扩展；

2. 针对大语言模型增量推理阶段kv cache内存占用高和搬移开销大的问题，提出基于语义

信息的key聚类value融合的kv cache高压缩比算法，并设计推理框架层面的内存管理、压缩融合算子优化等方案，推理句长32K时kv cache压缩8倍，提升端到端推理速度2倍；

3. 为了保证推理的精度，结合Attention的稀疏性特征，提出hashtopk的以查代算的方法，将query和key cache通过局部敏感hash映射到低维空间，再做hamming距离计算得到topk索引，以该索引查询Attention计算时的重要kv，来降低kv cache的加载量，原始kv cache依然保存在GPU global memory中。该方法提升大语言模型增量推理80+%的吞吐。

相关研究结果形成6篇专利，1篇技术秘密，以及2篇在投论文，并且已上线华为MindIE等产品，以提升基于昇腾平台的大模型推理速度。

2021.7 – 2023.12　AI大模型训练加速方案研究

承担角色： 项目经理，团队5+博士生，设计并实现精度无损的存网算协同的训练加速方案

问题描述： 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。大模型参数量超过千亿级别，训练面临着严重的内存墙、效率墙、计算墙问题。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。扩展到分布式之后需要在内存使用量、通信数据量、计算效率之间权衡。

主要工作： 大模型训练需要成千上万张卡做计算，我们从内存、通信、计算等角度出发，在系统侧设计完全精度无损的方案，来提升训练效率：

1. 大模型分布式训练往往采用3D并行的策略，即数据并行(DP)、tensor并行(TP)、流水并行(PP)，3个并行维度上的通信因为计算依赖关系，均在critical path上，导致通信开销在端到端占比50%以上，硬件资源利用率(MFU)不足40%。针对该问题，我们提出3D并行维度上的通信隐藏方案，在DP维度以额外内存换计算与通信的解耦，TP维度细粒度切分计算和通信，PP维度前反向计算交叉排布，来增加3个维度上的通信与计算相互隐藏的机会，能够将端到端训练速度提升20%以上；

2. 大模型训练内存占用除权重外，一大比重来源于激活值(Activation)，前向计算产生的激活需要保留到反向计算时使用，因此业界提出重计算的方法，在反向使用激活时再重新计算，这会引入额外33%的计算开销来换取内存上的节省。但往往配置并行策略后，GPU物理内存并未完全用满，启发我们在内存占用和重算开销之间巧妙权衡，为此我们设计自适应重计算方案，选择内存占用高但计算开销低的操作优先做重算，训练速度至多提升30%；

3. 在流水并行的场景下，出现GPU内存利用率不均衡的现象，流水头部GPU往往内存占用更多。而又考虑到单服务器内部GPU之间带宽的异构性，我们提出D2D swap的技术来使得GPU内存负载均衡，从而能够训练更大的模型，并采用重计算、offload等先进技术，进一步提升单服务器的训练模型规模。

相关工作产生4篇专利，一篇HPCA论文，并且实际落地到华为公司训练平台MindSpeed，支撑基于昇腾平台的科大讯飞星火大模型、蚂蚁MoE大模型等模型的高效训练。

2018.1 – 2021.6　针对数据并行分布式深度学习系统的研究

承担角色： 主导完成问题探索、方案设计、系统实现和论文撰写工作。

问题描述： 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。数据并行策略是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。

主要工作： 基于数据并行的分布式深度学习训练面临这梯度同步制约训练过程的问题，如梯度同步能在整个训练过程中占90%以上。梯度压缩算法在几乎不影响训练精度的前提下，能够将梯度压缩甚至到原来的千分之一，为加速训练带来了一种可能，但梯度压缩算法在实际训练系统中对训练速度的提升非常有限，其原因在于：1. 梯度压缩算法与主流的梯度同步优化策略不兼容（如batching and partitioning），使得梯度同步过程步伐增加；2. 梯度压缩本身具有不可忽略的overhead，需要根据实际训练情况决策是否压缩；3. 梯度压缩算法的实际应用需要面向GPU的高效实现、系统注册等过程，专业门槛很高。针对以上原因，我们提出HiPress，一个高效、梯度压缩感知的数据并行深度学习训练框架，其主要包含三个组件：1. CaSync，利用pipelines技术隐藏梯度压缩相关的计算，利用bulk synchronization技术加速小任务的网络传输；2. SeCoPa，利用cost model分析梯度是否应该被压缩，以及切分成多少份；3. CompLL，总结梯度压缩算法的通用算子，并在GPU上做高效实现，进一步提出简单的Domain Specific Language和code generator，用户只需简单描述梯度压缩算法的逻辑，便可通过generator直接翻译成利用common operator组合实现的高效GPU代码。实验证明，HiPress比开源系统有1.2-15.4倍的训练速度提升。

相关研究和扩展发表在系统领域顶会SOSP 2021和IEEE TPDS上，并形成1篇专利。

获得奖励

工作阶段： 2024世界人工智能大会 青年优秀论文奖；
华为公司金牌团队奖，2023年12月；
大禹治水攻关行动特等奖，2025年1月；
华为难题揭榜火花奖一等奖两项，2023年12月，2024年9月；
中央研究院创新先锋一等奖两项，2023上半年、2024上半年；
公司年度工作会议奖-创新与技术突破奖，2023年12月；
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；2023年计算会战1030阶段；2024年计算会战530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。

论文成果

1. **Youhui Bai,** Cheng Li, Quan Zhou, Jun Yi, Ping Gong, Feng Yan, Ruichuan Chen and Yinlong Xu."Gradient Compression Supercharged High-Performance Data Parallel DNN Training", **USENIX SOSP (CCF推荐A类会议) 2021**
2. **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**
3. **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**
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", **GNNSys workshop 2021**
5. 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**
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. Zewen Jin, Sheng Wang, Jiaan Zhu, Hongrui Zhan, **Youhui Bai**, Lin Zhang, Zhenyu Ming, Cheng Li. "BigMac: A Communication-Efficient Mixture-of-Experts Model Structure for Fast Training and Inference", **AAAI (CCF推荐A类会议) 2025**
9. 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**
10. Shengnan Wang, **Youhui Bai**, Lin Zhang, Pingyi Zhou, Shixiong Zhao, Gong Zhang, Sen Wang, Renhai Chen, Hua Xu, Hongwei Sun. "XL3M: A Training-free Framework for LLM Length Extension Based on Segment-wise Inference", **arXiv 2024**

专利成果

1. 2023年3月，一种在昇腾芯片上基于cube单元加速向量类计算的方法。**白有辉**，王森，周华漫等。
2. 2023年6月，一种在内存复用和offload结合的场景下的内存消减和数据交换方法。**白有辉**，周华漫等。
3. 2023年7月，一种在昇腾芯片上加速attention计算并降低内存占用量的方法。**华为潜在高价值专利。**周华漫，**白有辉**等。
4. 2023年8月，一种神经网络数据并行训练场景下低秩压缩耦合的梯度同步方法。**白有辉**，王森等。
5. 2024年3月，一种在大模型增量推理场景下压缩kv cache降低内存占用的方法。王盛南，**白有辉**等。
6. 2024年5月，基于分段加权融合的大模型长序列推理框架。**技术秘密**。王盛南，**白有辉**等。
7. 2024年7月，一种在跨平台场景中屏蔽硬件差异降低算子开发门槛的方法。苏景波，**白有辉**等。
8. 2024年7月，一种基于分离式架构的Attention融合算子并行方法。**白有辉**，周华漫等。
9. 2025年2月，一种在大模型推理场景下利用hash函数加速attention计算的方法。王盛南，**白有辉**等。