The 21st ChinaSys Workshop

第 21 届 ChinaSys 研讨会

2021 年 12 月 4 日-5 日

中国 厦门

（线上会议）

会议程序手册

主办单位：ACM SIGOPS ChinaSys

承办单位：厦门大学、中科（厦门）数据智能研究院

第 21 届 ChinaSys 研讨会组织机构

主办单位

承办单位

赞助伙伴

ACM SIGOPS ChinaSys 介绍

ACM SIGOPS ChinaSys 是在 ChinaSys 的基础上成立的，是中国计算机系统研究者和从业者的共同体，旨在共享资源并为共同体成员提供交换思想和会晤的论坛。它关注的领域包括：操作系统、虚拟化技术、分布式系统和网络、系统安全、移动嵌入式系统、云计算、多核和众核系统，以及程序设计语言、计算机系统结构和计算机系统之间的交互，等等。

ChinaSys 研讨会的创立受到了欧洲计算机系统研讨会 (EuroSys) 和 WWW Computer Architecture 的启发。在 2011 年 7 月的第 2 届亚太区系统研讨会（APSys 2011, 上海）上，微软亚洲研究院张峥、上海交通大学陈海波召集来自清华大学、复旦大学、北京大学、华中科技大学、中国科学技术大学五所高校的老师以及来自微软亚洲研究院、中科院计算所的研究人员商讨成立 ChinaSys 组织，以促进中国在系统相关领域的研究和实践。同年 11 月，第一届 ChinaSys 研讨会在深圳正式举行。自成立以来，ChinaSys 每年举办组织两次仅限邀请对象出席的研讨会，以供参会人员交流探讨其早期研究成果或最近发表的高水平研究成果，迄今已成功举办 20 届。

2015 年 8 月，在 ACM SIGOPS 主席 Robbert van Renesse（美国康奈尔大学教授）和副主席 Shan Lu（芝加哥大学教授）的支持下，成立 ACM SIGOPS China Chapter – ChinaSys。2016 年冬，ChinaSys 组织成为 ACM China 的一部分。

ACM SIGOPS ChinaSys 组织机构

主席：陈海波（上海交通大学）

副主席：包云岗（中国科学院计算技术研究所）

秘书长：张　昱（中国科学技术大学）

ACM SIGOPS ChinaSys 奖励委员会成员

陈文光	清华大学
陈海波	上海交通大学
包云岗	中国科学院计算技术研究所
周礼栋	微软亚洲研究院
卢　山	美国芝加哥大学
华　宇	华中科技大学
张　昱	中国科学技术大学

ACM SIGOPS ChinaSys 执行委员会成员

陈文光	清华大学
陈海波	上海交通大学
包云岗	中国科学院计算技术研究所
张　昱	中国科学技术大学
臧斌宇	上海交通大学
张　峥	上海纽约大学
周礼栋	微软亚洲研究院
罗英伟	北京大学
廖小飞	华中科技大学
王　蕾	国防科技大学
张为华	复旦大学
喻之斌	中国科学院深圳先进技术研究院
计卫星	北京理工大学
毛　波	厦门大学

第 21 届 ChinaSys 研讨会

第 21 届 ChinaSys 研讨会由 ACM SIGOPS ChinaSys 主办，厦门大学和中科（厦门）数据智能研究院承办，华为、商汤科技和上海人工智能实验室赞助，将于 2021 年 12 月 4-5 日在厦门举行。参会者遍布世界各地，包括威斯康星麦迪逊分校、电子科大、清华大学、北京大学、中国科学技术大学、国防科大、上海交通大学、华东师范、南开大学、天津大学、厦门大学、乔治梅森大学、卡塔尔计算所、中科院计算所、中科院深圳先研院、香港大学、人大、华南师范、微软亚洲研究院等来自学术界的系统研究者，和商汤、华为、一流科技等来自工业界的系统开发者、使用者。本次研讨会将针对云计算、数据存储、神经网络、Serverless、系统安全、并行计算、可靠性、AI for Systems 等计算机系统领域展开深入交流。

第 21 届 ChinaSys 研讨会组织委员会

大会主席：舒继武（厦门大学/清华大学）

包云岗（中国科学院计算技术研究所）

本地主席：刘向荣（厦门大学）

徐勇军（中科（厦门）数据智能研究院）

程序委员会主席：张一鸣（厦门大学）

李　诚（中国科学技术大学）

第 21 届 ChinaSys 研讨会程序委员会

王　卅	中国科学院计算技术研究所
左鹏飞	华为
刘　珂	中国科学院计算技术研究所
刘盛云	上海交通大学
杜军朝	西安电子科技大学
李永坤	中国科学技术大学
李雨森	南开大学
李慧霸	阿里巴巴
沈志荣	厦门大学
杨　智	北京大学
张逢喆	华为
张　鹏	西安交通大学
陆游游	清华大学
陈　力	华为
陈汉华	华中科技大学
陈　全	上海交通大学
陈　浩	湖南大学
苗又山	微软亚洲研究院
郑　龙	华中科技大学
赵来平	天津大学
徐尔茨	国防科技大学
蒋炎岩	南京大学
翟季冬	清华大学

线上会议议程

12 月 4 日（周六）	
8:20-8:35	开幕式及颁奖仪式（优博、新星奖） 致辞嘉宾：舒继武（厦门大学/清华大学） 颁奖嘉宾：陈海波（上海交大）
Session：Keynote I Chair：李诚（中国科学技术大学）	
8:35-9:20	人工智能新时代对系统研究的机遇和挑战 林达华（香港中文大学&商汤联合创始人）
9:20-10:05	Parallelizing Single-Machine Graph Algorithms 樊文飞（爱丁堡大学、院士）
10:05-10:50	从随机中寻找秩序 马晓松（卡塔尔计算研究所主任科学家）
	Session 1: File and storage systems I Chair：沈志荣（厦门大学）
10:50-11:10	The Storage Hierarchy is Not a Hierarchy: Optimizing Caching on Modern Storage Devices with Orthus Kan Wu (University of Wisconsin-Madison)
11:10-11:30	Balancing Storage Efficiency and Data Confidentiality with Tunable Encrypted Deduplication Jingwei Li（电子科大）
11:30-11:50	ParaBit: Processing Parallel Bitwise Operations in NAND Flash Memory Based SSDs Congming Gao（清华大学）
11:50-13:00	午休
	Session 2: DL Frameworks Chair：赵来平（天津大学）
13:00-13:20	OneFlow: Redesign the Distributed Deep Learning Framework from Scratch Jinhui Yuan （一流科技）
13:20-13:40	Gradient Compression Supercharged High-Performance Data Parallel DNN Training Youhui Bai（中国科学技术大学）
13:40-14:00	Efficient sparse collective communication and its application to accelerate distributed deep learning Jiawei Fei (国防科技大学和阿卜杜拉国王科技大学)
	Session 3: AI for Systems Chair：杨智（北京大学）

14:00-14:20	DRLPart: A Deep Reinforcement Learning Framework for Optimally Efficient and Robust Resource Partitioning on Commodity Servers Ruobing Chen（南开大学）
14:20-14:40	ReTail: Opting for Learning Simplicity to Enable QoS-Aware Power Management in the Cloud Shuang Chen（华为）
14:40-15:00	Experiences of Landing Machine Learning onto Market-Scale Mobile Malware Detection Liangyi Gong (清华大学)
	Session 4: Matrix Computation Chair：蒋炎岩（南京大学）
15:00-15:20	Hybrid Evaluation for Distributed Iterative Matrix Computation Zihao Chen（华东师范）
15:20-15:40	Everything You Always Wanted to Know about Sparse Matrix-Vector/Matrix Multiplication on GPUs Guohao Dai（清华）
15:40-16:00	LibShalom: Optimizing Small and Irregular-Shaped Matrix Multiplications on ARMv8 Multi-Cores Weiling Yang（国防科大）
	Session 5: Microservice and Serverless Chair：陈全（上海交大）
16:00-16:20	Understanding, Predicting and Scheduling Serverless Workloads under Partial Interference Yanan Yang（天津大学）
16:20-16:40	Characterizing Microservice Dependency and Performance: Alibaba Trace Analysis Shutian Luo（深圳先研院）
16:40-17:00	FaaSNet: Scalable and Fast Provisioning of Custom Serverless Container Runtimes at Alibaba Cloud Function Compute Ao Wang（乔治梅森大学）
	Session 6: ML systems I Chair：左鹏飞（华为）
17:00-17:20	CHRONUS: A Novel Deadline-aware Scheduler for Deep Learning Training Jobs Wei Gao (商汤科技)
17:20-17:40	Distilling Bit-level Sparsity Parallelism for General Purpose Deep Learning Acceleration Hang Lu（计算所）
17:40-18:00	LongTail-Bench: A Benchmark Suite for Long-tail Operators in Deep Learning Xiuhong Li （商汤）

12 月 5 日（周日）	
Session：Keynote II Chair：张一鸣（厦门大学）	
8:45-9:30	面向高效能图神经网络的算法与体系结构协同设计 李肯立（湖南大学、长江学者、国家杰青）
9:30-10:15	AI 框架未来的挑战及 MindSpore 的实践 金雪锋（华为 MindSpore 首席架构师）
10:15-11:00	题目：异构高性能计算机编程模型和优化 翟季冬（清华大学、国家优青）
Panel Moderator：李诚（中国科学技术大学）	
11:00-12:00	Topic: ChinaSys 未来十年展望 Panelist：陈海波、包云岗、金雪锋、翟季冬、王肇国
12:00-13:00	午休
	Session 7: Reliability Chair：刘珂（中科院计算所）
13:00-13:20	Geometric Partitioning: Explore the Boundary of Optimal Erasure Code Repair Yingdi Shan (清华大学)
13:20-13:40	Boosting Full-Node Repair in Erasure-Coded Storage Shiyao Lin (厦门大学)
13:40-14:00	Crash Consistent Non-Volatile Memory Express Xiaojian Liao （清华大学）
	Session 8: ML Systems II Chair：李雨森（南开大学）
14:00-14:20	Tacker: Tensor-CUDA Core Kernel Fusion for Improving the GPU Utilization while Ensuring QoS Han Zhao（上海交大）
14:20-14:40	Elena: An End-to-end Compilation Framework for Freestyle Tensor Computing Codes in High-level Language Xiuhong Li （商汤）
14:40-15:00	HET: Scaling out Huge Embedding Model Training via Cache-enabled Distributed Framework Xupeng Miao （北京大学）
	Session 9: Testing and Security Chair：苗又山（微软亚洲研究院）
15:00-15:20	TwinVisor: Hardware-isolated Confidential Virtual Machines for ARM Dingji Li（上海交大）
15:20-15:40	BIDL: A High-throughput, Low-latency Permissioned Blockchain Framework for Datacenter Networks Ji Qi (The University of Hong Kong)

15:40-16:00	Accelerating Encrypted Deduplication via SGX Yanjing Ren（电子科大）
	Session 10: Testing and Understanding Chair：张鹏（西安交大）
16:00-16:20	An Evolutionary Study of Configuration Design and Implementation in Cloud Systems Yuanliang Zhang（国防科大）
16:20-16:40	Enable Simultaneous DNN Services Based on Deterministic Operator Overlap and Precise Latency Prediction Weihao Cui（上海交大）
16:40-17:00	Challenges and Opportunities: An In-depth Empirical Study on Configuration Error Injection Testing Wang Li （国防科大）
	Session 11: Potpourri Chair：王卅（中科院计算所）
17:00-17:20	G-TADOC: Enabling Efficient GPU-Based Text Analytics without Decompression Feng Zhang (中国人民大学)
17:20-17:40	Multi-Intention-Aware Configuration Selection for Performance Tuning Haochen He （国防科大）
17:40-18:00	DMA-assisted I/O for Persistent Memory Weijie Zhang (华南师范大学)
18:00-18:20	TensorSSA: Enabling Efficient Compilation Optimization for View-style Deep Learning Operations with MLIR Jinming Ma （商汤）
18:20	闭幕 致辞嘉宾：包云岗（中科院计算所）

特邀报告

Keynote 1

报告题目: 人工智能新时代对系统研究的机遇和挑战

摘要:

本报告将全面介绍商汤科技在 CNN 神经网络并行训练过程中进行的并行优化、编译优化、存储优化、网络优化、大模型优化等系统方面的探索和所面临的开放问题。

个人简介:

林达华,香港中文大学信息工程系副教授，商汤科技联合创始人，香港中文大学-商汤科技联合实验室主任，上海人工智能实验室教授。到香港中文大学任教前,他曾于 2012 年到 2014 年在芝加哥丰田科技研究院任研究助理教授。

林达华教授在计算机视觉,概率推断，与深度学习方面有广泛的研究经历，并在多个课题上取得突出成绩。他在 CVPR/ICCV/ECCV/NIPS/PAMI 等计算机视觉与机器学习顶级会议与期刊发表逾 150+篇论文。他在 2010 年获得机器学习领域最权威国际会议 NIPS 的最佳学生论文奖，并在 2009 年与 2011 年获得计算机视觉最高学术会议 ICCV 的杰出评审员奖。他曾指导香港中文大学的研究团队参加 ImageNet、ActivityNet、以及 MSCOCO 等计算机视觉领域的主要国际竞赛，获得多个冠军。此外，他也担任 CVPR，ECCV，BMVC，ACM Multimedia 等主要国际会议的领域主席，以及顶级国际期刊 IJCV 的编委。

Keynote 2

报告题目：

Parallelizing Single-Machine Graph Algorithms

摘要：

This talk tackles two issues in connection with parallel graph computations.(1) Is it possible to simplify parallel programming, from think parallel to think sequential? That is, we want a parallel system such that we can plug in sequential graph algorithms, and the system parallelizes computations across a cluster of machines, without degradation in performance or functionality of existing graph query engines. (2) Does there exist a parallel model that optimizes computation by adaptively switching between BSP (Bulk Synchronous Parallel) and AP (Asynchronous Parallel) models? That is, the model retains the advantages of BSP and AP, while it reduces stragglers and redundant stale computations inherent to BSP and AP. We answer both questions in the affirmative.

个人简介：

樊文飞院士，英国爱丁堡大学信息学院主任教授，中国科学院外籍院士， 英国皇家学会院士、 欧洲科学院院士、英国爱丁堡皇家学会院士、计算机协会会士（ACM Fellow)。深圳计算科学研究院首席科学家、北京大数据科学与脑机智能高精尖创新中心首席科学家(北航)、北京大学深圳研究生院南燕荣誉教授、清华大学杰出客座教授。毕业于北京大学（本科，硕士）和美国宾夕法尼亚大学（博士），任职爱丁堡大学前为美国贝尔实验室科学家。 曾获得英国皇家学会 Wolfson 研究成果奖 (2018)、欧洲研究委员会 ERC Advanced Fellowship (2015）、英国 Roger Needham 奖（2008）、中国长江学者 (2007), 海外杰出青年学者（2003）、美国 CAREER Award（2001), Elsevier 网络科学刊物年度最佳论文和最杰出作者奖（2002）以及数据管理四大国际顶级 理论与系统会议的时间检验奖和最佳论文奖: Alberto O.Mendelzon 时间检验奖/ACMPODS 十年最佳论文奖 (2010 和 2015), ACM SIGMOD (2017)、VLDB（2010）和 ICDE（2007）最佳论文奖。

Keynote 3

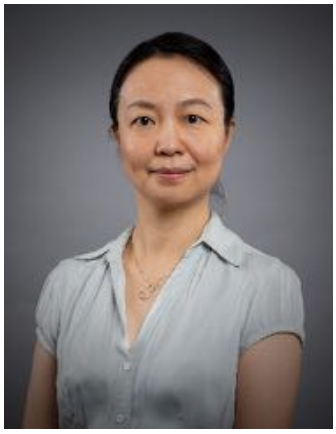
报告题目：

从随机中寻找秩序

摘要：

自然世界和人类社会中，都存在很多随机的现象，反映在计算应用里也是这样。有的属于数据访问上自然带有的随机性，有的属于处理上，现有系统通过随机方法来分布数据或访问，以达到性能优化目的，譬如负载均衡。然而，由于当前处理器和内存的架构设计，随机访问和顺序访问之间仍有较大的性能差距。我们将简要介绍近期的三个工作中，是怎样将图系统或分布式存储中大量的随机访问或数据放置，转换为顺序的或有序的存贮及使用，并且分享我们在研究过程中的观察和教训。

个人简介：



马晓松，卡塔尔计算研究所主任科学家，ACM 杰出会员。北京大学计算机系本科，伊利诺伊大学香槟分校博士。研究图计算、存储系统、分布式及云计算等等。曾获美国 NSF 事业奖，能源部事业奖，及 UIUC 计算机系杰出教育者校友奖。她目前担任 ACM Transactions on Storage 的编辑，在 SOSP, OSDI, FAST, EuroSys, ATC, NSDI 等主要系统会议上发表论文十余篇，并曾获 HPDC 20 年最佳论文奖，戈登贝尔奖提名，及两次 SC 会议最佳论文提名。

Keynote 4

报告题目: 面向高效能图神经网络的算法与体系结构协同设计

摘要：最近，图神经网络 (GNN) 将深度学习扩展到了面向图结构数据的学习，并在很多任务上展示了其强大的图表示学习能力。典型的图神经网络模型大都采用邻域消息传播机制，通过聚合邻居节点的特征来更新目标节点的特征。通过分析，我们发现邻域消息传播机制的简单实现会导致大量的冗余计算和冗余通信开销。本报告主要介绍我们在高效能图神经网络方面的一些研究工作。我们通过算法与体系结构的协同设计来去除计算过程中的冗余，以此来加速图神经网络。与目前的图神经网络加速器相比，我们提出的图神经网络加速器在不损失网络精度的情况下带来了可观的加速比并极大的降低了能耗。

个人简介

李肯立，湖南大学校长助理，信息科学与工程学院院长，国家超级计算长沙中心主任，教育部长江学者特聘教授、国家杰出青年科学基金获得者、中组部万人计划科技创新领军人才、科技部创新人才推进计划入选者。教育部高效能计算学科创新引智基地负责人、高性能计算应用软件教育部工程中心主任、数据分析湖南省工程技术研究中心主任。担任国家超级计算创新联盟副理事长、新一代人工智能产业技术创新联盟专家委员会委员、IEEE 高级会员、CCF 理事、CCF 杰出会员、CCF 高性能计算专业委员会常务委员等学术兼职，IEEE-TC 编委，IEEE-TII 等特刊编委。

Keynote 5

报告题目:

AI 框架未来的挑战及 MindSpore 的实践

摘要:

近些年来，机器学习/深度学习这类的应用负载对系统软件（特别是 AI 框架）提出了新的挑战，AI 框架如何与传统的编译器/编程语言和分布式并行技术结合，更好的支撑这些新的计算模式成为热点；本报告重点从 AI 框架的几大驱动力出发，包括 AI 模型、AI 的芯片以及 AI 的开发者体验，分析面向未来 AI 框架面临的挑战，并结合华为开源的 AI 框架 MindSpore 的实践，探讨技术演进之路。

个人简介:

金雪锋，MindSpore 首席架构师，华为 2012 实验室中央软件院架构与设计管理部部长，当前主要的方向 MLSys，包括大规模机器学习系统、AI 编译器、AI+科学计算系统等，之前曾先后担任华为的分布式数据库/大数据平台、电信基础软件平台的技术负责人，有 20 年系统软件设计和开发经验。

Keynote 6

报告题目：异构高性能计算机编程模型和优化

摘要：随着摩尔定律的逐渐放缓，大规模异构系统成为当前高性能计算领域发展的主流。庞大的系统规模以及复杂的体系结构导致在大规模异构高性能计算机上编写高效率的并行程序变得日益复杂。本报告主要讨论在当前以及下一代高性能计算机上，如何开展轻量级并行程序性能分析和优化，设计领域特定编程模型，降低用户在大规模异构高性能计算机上编写并行程序的复杂度，提高编程效率。

个人简介：

翟季冬，清华大学计算机系长聘副教授，博士生导师。现为清华大学计算机系高性能所副所长，ACM 中国高性能计算专家委员会秘书长、北京智源青年科学家。2015-2016 在斯坦福大学计算机系任访问助理教授。主要研究方向包括高性能计算、编程模型和编译系统等。研究成果发表在相关领域顶级学术会议和期刊——SC、ICS、PPOPP、ASPLOS、MICRO、OSDI、ATC、IEEE TC、IEEE TPDS 等。获 ICS 2021 最佳学生论文奖、SC 2014 Best Paper Finalist、ICDCS 2020 Best Paper Honorable Mention 奖。担任 NPC 2018 程序委员会主席、IEEE Cluster 2021 领域主席、SC 2022 领域副主席，SC、ICS、PPOPP、PACT 等国际学术会议程序委员会委员。目前担任《IEEE Transactions on Computers》、《IEEE Transactions on Parallel and Distributed Systems》、《IEEE Transactions on Cloud Computing》等多个国际学术期刊编委。担任清华大学学生超算团队教练，指导的团队十二次获得世界冠军。在 2015 年和 2018 年包揽了 SC、ISC、ASC 三大国际超算竞赛的总冠军，实现“大满贯”。获教育部科技进步一等奖、中国电子学会科学技术一等奖、中国计算机学会优秀博士学位论文奖、IEEE TPDS 杰出编委奖（Editorial Excellence Award）、国家自然科学基金优秀青年科学基金、CCF-IEEE CS 青年科学家奖。