# The 22nd ChinaSys Workshop

# 第 22 届 ChinaSys 研讨会

**2022 年 5 月 21 日-22 日**

线上

# 会议程序手册

主办单位：ACM China SIGOPS

承办单位：清华大学　北京理工大学　贵州财经大学

# 第 22 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。

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 21 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，即 ACM China SIGOPS。

# 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研究院 |
| 计卫星 | 北京理工大学 |
| 毛　波 | 厦门大学 |

# ACM SIGOPS ChinaSys 青年执行委员会成员

| | |
|---|---|
| 夏　文 | 哈尔滨工业大学（深圳） |
| 李博杰 | 华为技术有限公司 |
| 王肇国 | 上海交通大学 |
| 左鹏飞 | 华为技术有限公司 |
| 杜子东 | 中科院计算所 |
| 张　霁 | 华为技术有限公司 |
| 陈游旻 | 清华大学 |
| 李　诚 | 中国科学技术大学 |
| 魏星达 | 上海交通大学 |
| 姚　婷 | 华为技术有限公司 |

# 第 22 届 ChinaSys 研讨会

第 22 届 ChinaSys 研讨会将于 2022 年 5 月 21-22 日线上举行。本届研讨会由 ACM China SIGOPS (ChinaSys)主办，清华大学、北京理工大学和贵州财经大学承办，并由华为技术有限公司和蚂蚁集团赞助。本届研讨会特别邀请了来自哥伦比亚大学、中科院计算所、微软亚洲研究院、芝加哥大学、中科院信工所和华为技术有限公司的顶级专家对学术和业界前沿动态进行分享。

本次会议报告人既有来自清华大学、北京大学、上海交通大学、南京大学、浙江大学、国防科技大学、中国科学院、卡内基梅隆大学、香港大学等国内外高等院校的系统研究者，又有来自微软亚洲研究院、华为、阿里巴巴等企业的系统研发者，并针对云计算、数据存储、数据库系统、人工智能、系统软件、安全等计算机系统领域展开深入交流。

# 第 22 届 ChinaSys 研讨会组织委员会

**大会主席：** 陈文光（清华大学）

计卫星（北京理工大学）

**本地主席：** 邓明森（贵州财经大学）

**程序委员会主席：** 华　宇（华中科技大学）

王肇国（上海交通大学）

# 第 22 届 ChinaSys 研讨会程序委员会

（按姓氏拼音排序）

| | |
|---|---|
| 陈　康 | 清华大学 |
| 陈　全 | 上海交通大学 |
| 崔鹤鸣 | 香港大学 |
| 杜子东 | 中国科学院计算所 |
| 郭振宇 | 蚂蚁集团 |
| 何水兵 | 浙江大学 |
| Ryan Huang | Johns Hopkins University |
| Zhihao Jia | Carnegie Mellon University |
| 金　鑫 | 北京大学 |
| Jinyang Li | New York University |
| 刘海坤 | 华中科技大学 |
| 刘宇涛 | 华为技术有限公司 |
| 陆游游 | 清华大学 |
| Shuai Mu | Stony Brook University |
| 单一舟 | 华为技术有限公司 |
| 孙广宇 | 北京大学 |
| 孙园园 | 阿里巴巴达摩院 |
| Cheng Tan | Northeastern University |
| 王敏捷 | 亚马逊上海人工智能研究院 |
| 夏　文 | 哈尔滨工业大学（深圳） |
| 夏虞斌 | 上海交通大学 |

| | |
|---|---|
| Tianyin Xu | UIUC |
| Ding Yuan | University of Toronto |
| 翟季冬 | 清华大学 |
| 张权路 | 微软亚洲研究院 |
| Yongle Zhang | Purdue University |
| 张　源 | 复旦大学 |
| 赵永威 | 中国科学院计算所 |
| Zhijia Zhao | University of California, Riverside |
| 左鹏飞 | 华为技术有限公司 |

# 会议议程

| 5 月 21 日 | |
|---|---|
| 8:20-8:30 | 开幕 |
| **Keynote I**<br>(Session Chair: 华宇, 华中科技大学) | |
| 8:30-9:20 | Debugging Performance Issues in Modern Desktop Applications<br>Junfeng Yang（Columbia University） |
| 9:20-10:10 | 高性能计算性能工程<br>谭光明（中国科学院计算技术研究所） |
| 10:10-11:00 | Acoustic Sensing and Applications<br>Lili Qiu（Microsoft Research Asia and UT Austin） |
| **Memory and Cache**<br>(Session Chair: 杜子东, 中国科学院计算技术研究所) | |
| 11:00-11:20 | Segcache: A Memory-Efficient and Scalable In-Memory Key-Value Cache for Small Objects<br>杨骏骋（卡内基梅隆大学） |
| 11:20-11:40 | NVAlloc: Rethinking Heap Metadata Management in Persistent Memory Allocators<br>党政（浙江大学） |
| 11:40-12:00 | CacheSifter: Sifting Cache Files for Boosted Mobile Performance and Lifetime<br>梁宇（香港城市大学） |
| **午休（12:00 – 13:00）** | |
| **Cloud and DB**<br>(Session Chair: 夏文, 哈尔滨工业大学（深圳）) | |
| 13:00-13:20 | Plugsched: A Safe and Efficient Live Update Approach for Cloud OS Scheduler<br>马腾（阿里巴巴） |
| 13:20-13:40 | Fluid: Dataset Abstraction and Elastic Acceleration for Cloud-native Data-Intensive Applications<br>顾荣（南京大学） |
| 13:40-14:00 | LOCAT: Low-Overhead Online Configuration Auto-Tuning of Spark SQL Applications<br>辛锦瀚（中国科学院深圳先进技术研究院） |
| **Transactions**<br>(Session Chair: 单一舟，华为技术有限公司) | |
| 14:00-14:20 | Forerunner: Constraint-Based Speculative Transaction Execution for Ethereum<br>郭众鑫（微软亚洲研究院） |
| 14:20-14:40 | Aurogon: Taming Aborts in All Phases for Distributed In-Memory Transactions<br>姜天洋（清华大学） |
| **NVM and Storage**<br>(Session Chair: 糜泽羽，上海交通大学) | |
| 14:40-15:00 | HTMFS: Strong Consistency Comes for Free with Hardware Transactional Memory in Persistent Memory File Systems<br>董明凯（上海交通大学） |

| | |
|---|---|
| 15:00-15:20 | Clio: A Hardware-Software Co-Designed Disaggregated Memory System<br>单一舟（华为） |
| 15:20-15:40 | Hardware-Augmented Page Prefetching for Disaggregated Memory<br>李海锋（中国科学院计算技术研究所） |
| 15:40-16:00 | Separating Data via Block Invalidation Time Inference for Write Amplification Reduction in Log-Structured Storage<br>王秋平（香港中文大学） |
| 16:00-16:20 | $MT^2$: Memory Bandwidth Regulation on Hybrid NVM/DRAM Platforms<br>董明凯（上海交通大学） |
| **Bugs & Security**<br>(Session Chair: 张源，复旦大学) | |
| 16:20-16:40 | Understanding and Finding On-the-Fly Configuration Bugs<br>王腾（国防科技大学） |
| 16:40-17:00 | HyBP: Hybrid Isolation-Randomization Secure Branch Predictor<br>赵路坦（中国科学院信息工程研究所） |
| **Concurrency**<br>(Session Chair: 刘宇涛，华为技术有限公司) | |
| 17:00-17:20 | Asymmetry-Aware Scalable Locking<br>古金宇（上海交通大学） |
| 17:20-17:40 | TileSpGEMM: A Tiled Algorithm for Parallel Sparse General Matrix-Matrix Multiplication on GPU<br>牛宇瑶（中国石油大学） |

| 5 月 22 日 | |
|---|---|
| **Keynote II**<br>(Session Chair: 王肇国， 上海交通大学) | |
| 8:30-9:20 | 15 Years of Learning From Mistakes in Building System Software<br>Shan Lu（The University of Chicago） |
| 9:20-10:10 | 处理器安全分支预测器<br>侯锐（中国科学院信息工程研究所） |
| 10:10-11:00 | 基于 DPU 的存储创新<br>黄克骥（华为） |
| **Architecture for Systems （I）**<br>(Session Chair: 沈志荣，厦门大学) | |
| 11:00-11:20 | Rolis: A Software Approach to Efficiently Replicating Multi-Core Transactions<br>沈维海（纽约州立大学石溪分校） |
| 11:20-11:40 | Romou: Rapidly Generate High-Performance Tensor Kernels for Mobile GPUs<br>梁任冬（加利福尼亚大学尔湾分校） |
| 11:40-12:00 | Optimized MPI Collective Algorithms for Dragonfly Topology<br>丰光南（中山大学） |
| **午休（12:00-13:00）** | |
| **Architecture for Systems（II）**<br>(Session Chair: 吴明瑜，上海交通大学) | |
| 13:00-13:20 | Sanger: A Co-Design Framework for Enabling Sparse Attention Using Reconfigurable Architecture<br>卢丽强（北京大学） |
| 13:20-13:40 | 基于 RISC-V 的用户态虚拟机监控器 DuVisor<br>糜泽羽 (上海交通大学) |
| **Software Services**<br>(Session Chair: 左鹏飞，华为技术有限公司) | |
| 13:40-14:00 | FaaSFlow: Enable Efficient Workflow Execution for Function-as-a-Service<br>李子俊（上海交通大学） |
| 14:00-14:20 | Web 服务的自动服务器无感知卸载技术<br>吴明瑜（上海交通大学） |
| 14:20-14:40 | Semantics Foundation for Cyber-Physical Systems Using Higher-Order UTP<br>徐雄（中国科学院软件研究所） |
| 14:40-15:00 | INFless: A Native Serverless System for Low-Latency, High-Throughput Inference<br>杨亚南（天津大学） |
| **System for ML**<br>(Session Chair: 董明凯， 上海交通大学) | |
| 15:00-15:20 | NeutronStar: Distributed GNN Training with Hybrid Dependency Management<br>王千阁（东北大学） |
| 15:20-15:40 | VELTAIR: Towards High-Performance Multi-Tenant Deep Learning Services via Adaptive Compilation and Scheduling<br>刘子汉 （上海交通大学） |
| 15:40-16:00 | NASPipe: High Performance and Reproducible Pipeline Parallel Supernet Training via Causal Synchronous Parallelism<br>赵世雄（香港大学） |

| | |
|---|---|
| 16:00-16:20 | HET-GMP: A Graph-based System Approach to Scaling Large Embedding Model Training<br>苗旭鹏（北京大学） |
| 16:20-16:40 | FasterMoE: Modeling and Optimizing Training of Large-Scale Dynamic Pre-Trained Models<br>何家傲（清华大学） |
| **Compression**<br>(Session Chair: 张权路，微软亚洲研究院) | |
| 16:40-17:00 | CompressDB: Enabling Efficient Compressed Data Direct Processing for Various Databases<br>张峰（中国人民大学） |
| 17:00-17:20 | TVStore: Automatically Bounding Time Series Storage via Time-Varying Compression<br>安彦哲（清华大学） |
| 17:20-17:25 | 闭幕 |

# 特邀报告

（按照报告时间顺序）

## Keynote 1

报告题目：

## Debugging Performance Issues in Modern Desktop Applications


摘要：

Modern desktop applications involve many asynchronous, concurrent interactions that make performance issues difficult to diagnose. Although prior work has used causal tracing for debugging performance issues in distributed systems, we find that these techniques suffer from high inaccuracies for desktop applications. In this talk, I will present Argus, a fast, effective causal tracing tool for debugging performance anomalies in desktop applications. Argus introduces a novel notion of strong and weak edges to explicitly model and annotate trace graph ambiguities, a new beam-search-based diagnosis algorithm to select the most likely causal paths in the presence of ambiguities, and a new way to compare causal paths across normal and abnormal executions. We have implemented Argus across multiple versions of macOS and evaluated it on 12 infamous spinning pinwheel issues in popular macOS applications. Argus diagnosed the root causes for all issues, 10 of which were previously unknown, some of which have been open for several years. This work won a Best Paper award in USENIX ATC 2021. It is joint with Lingmei Weng (lead PhD student, graduating next academic year), Ryan Peng Huang, and Jason Nieh.


个人简介：

Junfeng Yang is Professor of Computer Science, Member of the Data Science Institute, and co-Director of the Software Systems Lab at Columbia University. Yang's research centers on building reliable, secure, and fast software systems. Today's software systems are large, complex, and plagued with errors, some of which have caused critical system failures, breaches, and performance degradation. Yang has invented techniques, algorithms,

and tools to analyze, test, debug, monitor, and optimize real-world software, including Android, Linux, production systems at Microsoft, machine learning systems, and self-driving platforms, benefiting hundreds of millions of users. His research has resulted in numerous vulnerability patches to real-world systems, practical adoption at the largest technology companies, and press coverage at Scientific American, The Atlantic, The Register, Communications of ACM, and other news outlets.

Yang received BS in Computer Science from Tsinghua University and MS and PhD in Computer Science from Stanford University. He won the Sloan Research Fellowship and the Air Force Office of Scientific Research Young Investigator Program Award, both in 2012; the National Science Foundation CAREER award in 2011; the inaugural Rock Star Award of the Association of Chinese Scholars in Computing in 2019; and Best Paper Awards at the USENIX Symposium on Operating System Design and Implementation in 2004, the ACM Symposium on Operating Systems Principles in 2017, and the USENIX Annual Technical Conference in 2021.

## Keynote 2

报告题目:

# 高性能计算性能工程


摘要:

高性能计算领域的核心命题是关于如何满足应用性能需求，与一般性计算问题相比而言，性能通常是第一优先级考虑的指标。总体上而言，影响性能的诸多因素主要包括：硬件设计（流水线、向量宽度、Cache 大小等）、算法模型（复杂度等）、实现方式（编程语言、数据结构、库的版本等）、代码生成（编译器）、系统配置（操作系统的选择等）和执行环境（亲和性选择、资源分配和系统噪音等）。在真实的运行系统中，这些性能因素之间不是独立正交，而是相互影响形成一个非常复杂庞大的优化空间。在单纯以软件工程驱动的高性能计算软件栈设计中，人们为了追求高的生产效率，通过分层模块设计把错综复杂的性能因素“粗暴”地割裂开，在通用硬件性能提升放缓的情况下，所谓的软件“肿胀”导致的性能瓶颈问题就凸显出来。这种性能损失对以性能为第一优先目标的高性能计算而言显得尤为突出，因此，在继高性能计算的硬件工程和软件工程技术系统发展多年之后，本报告试图提倡高性能计算性能工程的研究，以系统发展性能工程技术，应对高性能计算软硬件栈在后摩尔时代的挑战。


个人简介:


谭光明，研究员、博导、中科院计算技术研究所高性能计算机研究中心主任。国家杰出青年基金获得者，参与了曙光系列高性能计算机包括曙光 4000/5000/6000/7000 系统研制。发表学术论文 100 余篇，包括 CCF A 类论文（TC、SC、PPoPP）和 Nature 子刊等，曾任 IEEE TPDS 编委和国际会议（SC、PPoPP）等程序委员。曾获得国家科技进步奖二等奖、卢嘉锡青年人才奖和全国向上向善好青年称号。

## Keynote 3

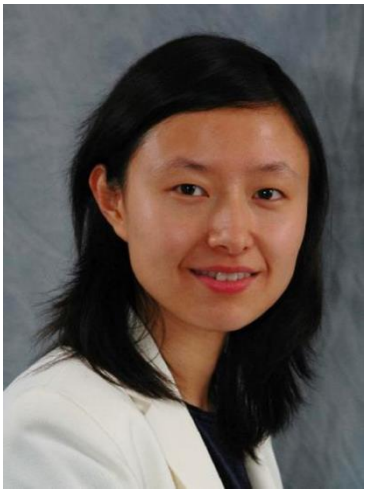
报告题目：

## Acoustic Sensing and Applications

摘要：

Video games, Virtual Reality (VR), Augmented Reality (AR), and Smart appliances (e.g., smart TVs and drones) all call for a new way for users to interact and control them. Motivated by this observation, we have developed a series of novel acoustic sensing technologies by transmitting specifically designed signals or using signals naturally arising from the environments. We further develop a few interesting applications on top of our motion tracking technology such as a follow-me drone and acoustic imaging on mobile phones.

个人简介：



Lili Qiu is an Assistant Managing Director at Microsoft Research Asia and a Professor at Computer Science Dept. in UT Austin. She got M.S. and PhD degrees in Computer Science from Cornell University in 1999 and 2001, respectively. After graduation, she spent 2001-2004 as a researcher at System & Networking Group in Microsoft Research Redmond. She joined UT Austin in 2005, and has founded a vibrant research group working on Internet and wireless networks at UT. She is an ACM Fellow and IEEE Fellow. She also got an NSF CAREER award and Google Faculty Research Award, and best paper awards at ACM MobiSys'18 and IEEE ICNP'17. She advised a PhD dissertation that won SIGMOBILE best dissertation award in 2020.

## Keynote 4

报告题目:

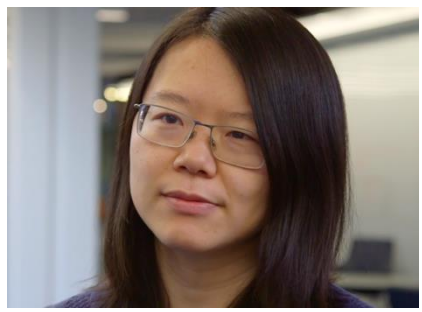
## 15 Years of Learning from Mistakes in Building System Software


摘要:

Bugs severely threaten the correctness and efficiency of software. With our system software growing its complexity, bugs in system software also evolve, imposing different challenges over the years.

In this talk, we look back at our study of concurrency bugs in multi-threaded software, which was done 15 years ago and recently won ASPLOS Influential Paper Award, as well as various bug studies that we conducted over the years about distributed systems, industry cloud systems, database systems, machine learning systems, etc. We discuss the lessons that we have learned, as well as the new challenges faced by today's system building.


个人简介:



Shan Lu is a Professor in the Department of Computer Science at the University of Chicago. Her research focuses on detecting, diagnosing, and fixing functional and performance bugs in software systems.

Shan is an ACM Distinguished Member (2019 class) and an Alfred P. Sloan Research Fellow (2014). Her co-authored papers have won distinguished paper and influential paper awards at ASPLOS, SOSP, OSDI, FAST, ICSE, FSE, CHI, and PLDI. Shan currently serves as the Chair of ACM-SIGOPS, and the Vice Chair of ACM SIG Governing Board Executive Committee. She served as the technical program co-chair for ASPLOS 2022, OSDI 2020, APSys 2018, and USENIX ATC 2015.

## Keynote 5

报告题目：

# 处理器安全分支预测器


摘要:

在数字化日益普及的今日，数据中心处理器芯片安全问题愈发重要。尤其是在云端，处理器面临着众多的安全风险。我们以处理器中性能提升关键模块——分支预测器——为切入点，分别从更新策略、内容存储、索引映射三个方面对传统分支预测器设计进行了解构，并提出了一系列安全增强机制，实现了分支预测器的安全重构。


个人简介:

侯锐，中国科学院信息工程研究所信息安全国家重点实验室副主任、研究员、博士生导师，获得国家杰青、优青项目资助。长期从事处理器芯片架构设计及芯片安全等方面的研究工作。

## Keynote 6

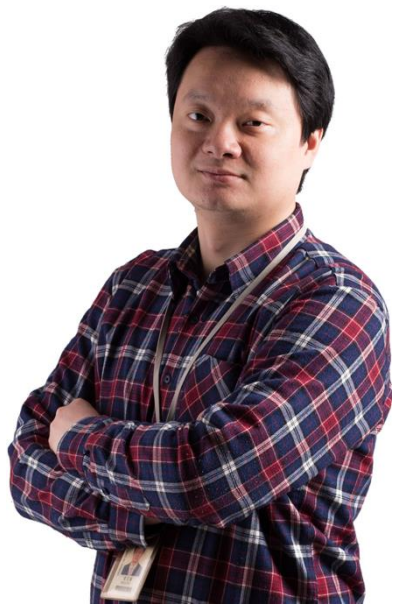
报告题目：

# 基于 DPU 的存储创新


摘要：

数据密集型业务的快速发展驱动 DPU 成为继 CPU 和 GPU 之后的第三大算力，业界围绕 DPU 进行创新也成为新的热点。华为依托在存储领域多年的技术积累，也基于 DPU 进行了从硬件和 OS 到虚拟化、大数据、数据库等场景加速的一些创新实践，旨在通过 DPU 与存储结合为用户带来数据处理和存储效率的倍数级提升。


个人简介：



黄克骥博士，华为存储领域 8 级技术专家，数据存储产品线首席架构师，负责华为存储产品的整体架构规划和技术演进，是华为存储产品竞争力实现业界领先的领军人物。黄克骥博士具有超过 18 年 ICT 从业经验和超过 15 年存储领域研究经验，持续深耕技术创新和根科技构建，在存储领域积累了深厚的技术功底，先后负责华为赛门铁克云存储产品、华为第一代分布式 NAS 产品和大数据存储产品、融合存储 NAS 产品、全闪存存储产品、华为第一代存储平台、华为云存储和数据架构等多个产品及大型架构的规划和设计，奠定了华为存储产品竞争力成为国内第一并进入 Gartner 领导者象限的坚实基础。