The 19th ChinaSys Workshop

第 19 届 ChinaSys 研讨会

2020 年 12 月 5 日-6 日

中国 重庆

会议程序册

主办单位：ACM China SIGOPS

承办单位：中国科学院计算技术研究所

中国计算机学会分布式计算与系统专委会

第 19 届 ChinaSys 研讨会组织机构

主办单位

承办单位

中国科学院计算技术研究所

中国计算机学会分布式计算与系统专委会

第 19 届 ChinaSys 组织委员会

大会主席：谭光明（中国科学院计算技术研究所）
大会主席：包云岗（中国科学院计算技术研究所）
程序委员会主席：陆游游（清华大学）
程序委员会主席：刘珂　（中国科学院计算技术研究所）

第 19 届 ChinaSys 程序委员会

（按姓氏拼音排序）

姓名	单位
陈泉	上海交通大学
陈超	中科院深圳先进技术研究院
杜子东	中国科学院计算技术研究所
高婉玲	中国科学院计算技术研究所
韩睿	中国科学院计算技术研究所
黄群	北京大学
姜艳艳	南京大学
李景伟	中国电子科技大学
李永坤	中国科学技术大学
李力	中国科学技术大学
刘晓东	国防科技大学
糜泽羽	上海交通大学
沈志荣	厦门大学
王卅	中国科学院计算研究所
夏文	哈尔滨工业大学（深圳）
范扬	微软
杨志	北京大学
尹舒	上海科技大学
翟继东	清华大学
郑龙	华中科技大学
郑佳琦	南京大学

会议日程

第一天 （ 12 月 5 日）	
08:30 – 08:50	**Opening Remarks**
Keynote Session 1	**Chair: 包云岗**
08:50 - 09:20	**高性能固态存储系统的构建** Speaker: 舒继武
09:20 - 09:50	**Project Organon: Building the Next Generation AI Software Infrastructure** Speaker: 张霖涛
09:50 – 10:00	茶歇
Keynote Session 2	**Chair: 谭光明**
10:00 - 10:30	**Tackling synchronization challenges in large scale industry code** Speaker: 卢山
10:30 - 11:00	**AI 工程的系统优化在阿里云 PAI 平台的实践** Speaker: 林伟
Paper Session 1: OSDI	
11:00 - 11:20	**Rammer: Enabling Holistic Deep Learning Compiler Optimizations with rTasks** Lingxiao Ma (北京大学&MSRA), Zhiqiang Xie (上海科技大学&MSRA), Zhi Yang (北京大学), Jilong Xue, Youshan Miao, Wei Cui, Wenxiang Hu, Fan Yang, Lintao Zhang, Lidong Zhou (MSRA)
11:20 - 11:40	**Write Dependency Disentanglement with HORAE** Xiaojian Liao, Youyou Lu, Erci Xu and Jiwu Shu (清华大学)
11:40 - 12:00	**Retiarii: A Deep Learning Exploratory-Training Framework** Quanlu Zhang, Zhenhua Han, Fan Yang, Yuge Zhang, Zhe Liu, Mao Yang and Lidong Zhou (MSRA)
12:00 – 1:30	午餐
Paper Session 2: Storage	
13:30 - 13:50	**Spool: Reliable Virtualized NVMe Storage Pool in Public Cloud Infrastructure** Shuai Xue, Shang Zhao, Quan Chen (上海交通大学), Gang Deng, Zheng Liu, Jie Zhang, Zhuo Song, Tao Ma, Yong Yang, Yanbo Zhou, Keqiang Niu, Sijie Sun (阿里云), Minyi Guo (上海交通大学)
13:50 - 14:10	**HDDse: Enabling High-Dimensional Disk State Embedding for Generic Failure Detection System of Heterogeneous Disks in Large Data Centers** Ji Zhang (华中科技大学)

14:10 - 14:30	**Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching** Yaohua Wang (国防科技大学)
Paper Session 3: Performance and Reliability	
14:30 - 14:50	**Effective Detection of Sleep-in-atomic-context Bugs in the Linux Kernel** Jia-Ju Bai (清华大学), Julia Lawall (INRIA), Shi-Min Hu (清华大学)
14:50 - 15:10	**Alita: Comprehensive Performance Isolation through Bias Resource Management for Public Clouds** Quan Chen, Shuai Xue, Shang Zhao (上海交通大学), Shanpei Chen, Yihao Wu, Yu Xu, Zhuo Song, Tao Ma, Yong Yang (阿里云), Minyi Guo (上海交通大学)
15:10 - 15:30	**Characterizing Serverless Platforms with ServerlessBench** Tianyi Yu, Qingyuan Liu, Dong Du, Yubin Xia, Binyu Zang and Haibo Chen (上海交通大学)
15:30 – 15:40	茶歇
Paper Session 4: Security	
15:40 - 16:00	**ZeroSpy: Exploring Software Inefficiency with Redundant Zeros** Xin You, Hailong Yang, Zhongzhi Luan, Depei Qian (北京航空航天大学), Xu Liu (北卡罗来纳州立大学)
16:00 - 16:20	**PDiff: Semantic-based Patch Presence Testing for Downstream Kernels** Zheyue Jiang, Yuan Zhang (复旦大学), Jun Xu (史蒂文斯理工学院), Qi Wen, Zhenghe Wang, Xiaohan Zhang (复旦大学), Xinyu Xing (宾夕法尼亚州立大学), Min Yang and Zhemin Yang (复旦大学)
16:20 - 16:40	**Detecting Kernel Refcount Bugs with Two-Dimensional Consistency Checking** Xin Tan, Yuan Zhang, Xiyu Yang (复旦大学), Kangjie Lu (明尼苏达大学), Min Yang (复旦大学)
Paper Session 5: Application	
16:40 - 17:00	**A Locality-Aware Energy-Efficient Accelerator for Graph Mining Applications** Pengcheng Yao, Long Zheng, Zhen Zeng, Yu Huang, Chuangyi Gui, Xiaofei Liao, Hai Jin and Jingling Xue (华中科技大学)
17:00 - 17:20	**BORA: A Bag Optimizer for Robotic Analysis** Jian Zhang and Shu Yin (上海科技大学)
17:20 - 17:40	**Ptolemy: Architecture Support for Robust Deep Learning** Yuxian Qiu, Jingwen Leng (上海交通大学), Yiming Gan (罗切斯特大学), Minyi Guo (上海交通大学), Yuhao Zhu (罗切斯特大学)
	晚宴

第二天 （ 12 月 6 日）	
Keynote Session 3 **Chair: 陆游游**	
09:00 - 09:30	**计算机系统研究的一些体会** Speaker: 陈海波
09:30 – 10:00	**HPC + AI + 物理模型：智能超算应用的新发展** Speaker: 贾伟乐
10:00 – 10:10	茶歇
Special Session: 科研背后的故事 **Chair: 李诚**	
10:10 - 10:30	**新星奖获得者 1**
10:30 - 10:50	**新星奖获得者 2**
10:50 - 11:10	**优博奖获得者 1**
11:10 - 11:30	**优博奖获得者 2**
11:30 - 12:00	**Panel** 陈海波，新星奖获得者 1，新星奖获得者 2，优博奖获得者 1 和优博奖获得者 2

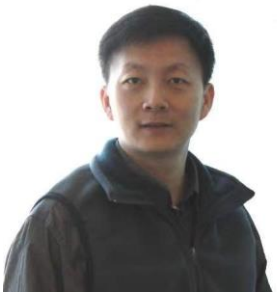
特邀报告

Keynote 1 舒继武

题目：高性能固态存储系统的构建

摘要：随着固态存储（Solid-State Drive, SSD）技术的不断革新，硬件性能有极大的提升，但系统软件开销在整个存储系统中所占比例越来越重。因此，设计高效的系统软件是构建高性能固态存储系统的关键问题。同时，CPU 多核架构的演进、高性能固态存储（NVMe SSD）的新硬件特性，为软硬件协同设计带来了机遇和挑战。本报告介绍固态存储上两个新的文件系统设计，以充分利用硬件的性能和特性：(1) 多核并发的闪存文件系统，提出新的文件系统并发控制机制、并发且闪存友好的数据索引以及高并发强一致的多流持久化技术，提升文件系统充分利用固态存储高带宽的能力；(2) 软硬件协同的双接口文件系统，利用 SSD 中的一小块非易失内存，在传统块存储栈以外构建独立的 I/O 访问路径，以加速文件系统的非对齐小写。

简介：舒继武，厦门大学信息学院院长，清华大学计算机系长聘教授，教育部长江学者特聘教授，国家杰出青年基金获得者，IEEE Fellow，中国计算机学会会士、信息存储技术专业委员会主任，灾备技术国家工程实验室副主任；国际期刊《ACMTransactions on Storage 》和《 IEEE Transactions onComputers》的 Associate Editor；主要研究领域为新型 NVM 存储系统与技术、闪存存储系统与技术、网络（/云/大数据）存储系统、存储可靠性与安全等，相关成果发表在包括 FAST、OSDI、USENIX ATC、MICRO、ISCA、ASPLOS、SC、EuroSys、DAC 等重要国际学术会议和 IEEE/ACM Trans.等权威期刊上；承担过国家重点研发项目、863 项目和课题、973 课题、国家自然科学基金重点项目等。获国家科技进步二等奖和国家发明技术二等奖各 1 次。



Keynote 2 张霖涛

Topic: Project Organon: Building the Next Generation AI Software Infrastructure

Abstract: In this talk, we will discuss Project Organon, an umbrella project in Microsoft Research Asia that aims at building the next generation software infrastructure for AI workloads. We noticed that currently the most widely used AI software stacks were usually developed by many separate and un-coordinated teams. These stacks often lack a coherent design, and often cannot perform holistic optimizations for the typical AI workloads. In project Organon, we aim to address this issue by designing a next-generation full-stack solution for AI workloads. We will describe the components of Project Organon and the approach we took to minimize the complexity of this project. Most parts of the project are open-sourced, and we welcome comments, collaborations, and contributions.

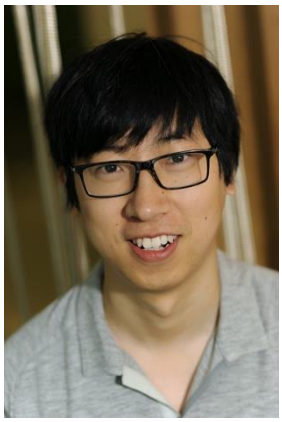
Speaker Bio: Dr. Lintao Zhang was a Partner Research Manager in the Systems and Networking Area of Microsoft Research Asia. Over his research career, Dr. Zhang has worked on a broad set of projects, including verification and logic, internet security, distributed systems, computer architecture, reconfigurable computing, and most recently datacenter software infrastructures for cloud-scale storage, networking, and computing. He has won many awards for his research, including award papers in SOSP and DATE, a 10-year retrospective most influential paper in ICCAD, the most cited paper in the 50-year history of DAC, a CAV Award, and a Richard Newton Technical Impact Award in Electronic Design Automation. He earned his Bachelor's degree from Peking University and his Ph.D. from Princeton University.

Keynote 3 卢山

Topic: Tackling synchronization challenges in large scale industry code

Abstract: Synchronizations are fundamental to the correctness and performance of concurrent software, by determining which operations can execute concurrently and which can-not. Unfortunately, correctly identifying all synchronizations has become extremely difficult in modern software systems due to the various forms of concurrency and various types of synchronization mechanisms. This talk presents our adventure in overcoming this challenge while analyzing Microsoft Azure system. In collaboration with Microsoft researchers, we have developed an unsupervised inference tool, called SherLock, to automatically identify synchronization operations (to appear in ASPLOS'21) and a fault-injection tool that has exposed more than 1000 synchronization bugs in Azure code without any synchronization annotation (SOSP'19 Best Paper).

Speaker Bio: Shan Lu is a Professor in the Department of Computer Science at the University of Chicago. Her research focuses on software reliability and efficiency. Shan is an ACM Distinguished Member (2019 class), an Alfred P. Sloan Research Fellow (2014), a Distinguished Educator Alumnus from Department of Computer Science at University of Illinois (2013), and NSF Career Award recipient (2010). Her co-authored papers won Best Paper Awards at ACM-SIGOPS SOSP 2019, USENIX OSDI 2016 and USENIX FAST 2013, 3 ACM-SIGSOFT Distinguished Paper Awards at ICSE 2019, ICSE 2015 and FSE 2014, an ACM-SIGPLAN Research Highlight Award at PLDI 2011, an IEEE Micro Top Picks in ASPLOS 2006, and a Google Scholar Classic Paper 2017. Shan currently serves as the Chair of ACM-SIGOPS (2019 --), Member-at-Large of ACM SIG Governing Board Executive Committee (2020 -- 2022), and the Associate Editor for IEEE Computer Architecture Letters. She served as the technical program co-chair for USENIX Symposium on Operating Systems Design and Implementation (OSDI) in 2020, USENIX Annual Technical Conference (ATC) in 2015, and ACM Asia-Pacific Systems Workshop (APSys) in 2018.



Keynote 4 林伟

题目：AI 工程的系统优化在阿里云 PAI 平台的实践

摘要：机器学习 PAI（Platform of Artificial Intelligence）是阿里云人工智能平台，提供一站式的机器学习解决方案。PAI 起初是服务于阿里巴巴集团内部（例如淘宝、支付宝和高德）的机器学习平台，致力于让公司内部开发者更高效、简洁、标准地使用人工智能 AI（Artificial Intelligence）技术。随着 PAI 的不断发展，2018 年 PAI 平台正式商业化，目前已经积累了数万的企业客户和个人开发者，是国内领先的云端机器学习平台之一。

近几年，深度学习算法在 GPU 的算力加持下证明了自己在各种领域内的巨大潜力。如今深度神经网络已经应用在阿里巴巴众多的产品中，横跨多个领域，包括计算机视觉、自然语言处理、语音识别，当然也包括对于阿里巴巴经济体非常重要的推荐和广告等。因此，深度学习已经成为阿里巴巴业务产品数据流中至关重要的一环。为了支持大规模的深度学习应用，阿里云 PAI 平台构建了大规模多租户共享的 GPU 集群，用来支持业务的发展。本次报告主要介绍阿里云 PAI 平台在大规模异构集群中如何通过系统调度，编译优化，分布式等系统工作提高用户开发模型的迭代效率。报告内容围绕 PAI 平台在 OSDI’ 20 会议上发表的两篇论文展开。

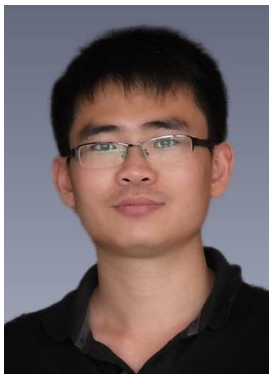
简介：林伟，阿里云智能研究员，阿里云机器学习 PAI 平台技术负责人，主攻大规模分布式训练加速、编译优化等 AI 工程的建设和性能优化。具有大规模并发系统有 15 年的系统架构设计及研发经验，并在国际一流 ODSI、NSDI、SIGMOD 会议上多次发表论文。原微软大数据平台组的核心成员，曾在微软亚洲研究院和微软美国工作 10 年，一直从事分布式系统开发和大数据平台的相关工作。

Keynote 5 陈海波

题目：计算机系统研究的一些体会

摘要：计算机系统对下管理硬件资源，对上提供应用执行环境，支撑着云计算、大数据、AI、物联网等新型计算模式的运行，因此计算机系统的研究也存在着独特的要求与挑战。该报告将介绍报告人在计算机系统研究中的一些体会，包括如何选择研究课题、如何参与国际学术界社区活动、如何形成学术品牌以及如何开展成果转化等。

简介：陈海波，上海交通大学特聘教授，并行与分布式系统研究所所长，领域操作系统教育部工程研究中心主任，国家杰出青年基金获得者、ACM 杰出科学家。主要研究领域为操作系统和系统安全。曾获教育部技术发明一等奖、中国青年科技奖、上海交通大学校长奖、CCF 青年科学家奖、全国优秀博士学位论文奖等。目前担任 ACM ChinaSys 主席、CCF 系统软件专委会副主任、ACM 旗舰杂志《Communications of the ACM》中国首位编委与 Special Sections 领域共同主席、《ACM Transactions on Storage》编委。曾任 ACM SOSP 2017 年大会共同主席、ACM CCS 2018 系统安全领域主席、ACM SIGSAC 奖励委员会委员。按照 csrankings.org 的统计，其在操作系统领域近 5 年（2015-2019）发表的高水平会议 (SOSP/OSDI, EuroSys, Usenix ATC 和 FAST）论文数居世界第一。



Keynote 6 贾伟乐

题目：HPC + AI + 物理模型：智能超算应用的新发展

摘要：智能超算为超算应用的发展提出了新的挑战和方向。如何融合传统的”HPC+物理模型”的计算模式与新的智能超算成为新的课题。本报告从典型的科学计算出发，以第一性原理分子动力学为例，展示一种全新的”HPC + AI + 物理模型”的计算模式。该计算模式以智能超算为硬件基础，成功将人工智能算法与物理模型数据结合。通过对算法的优化调优，我们在 Summit 超级计算机上首次实现了上亿原子的第一性原理分子动力学模拟，计算速度达到 1 纳秒/天。这比其他任何已知的第一性原理分子动力学模拟体系至少大 100 倍，计算速度至少快1000倍。最终DeePMD-kit在Summit全机上达到双精度91PFLOPS，混合单精度162PFLOPS，混合半精度 274PFLOPS。我们的分析显示：虽然 DeePMD-kit 达到了很好的性能，它仍是访存受限的应用，在混合半精度情况下尤其如此。这也为未来硬件设计提供了新的思路。

简介：贾伟乐，中国科学院计算技术研究所副研究员，入选中科院百人计划，SC20 戈登贝尔奖获奖人（一作）。2016 年博士毕业于中国科学院大学（计算机网络信息中心），之后加入加州大学伯克利分校从事博士后研究工作，主要研究方向为高性能计算、第一性原理计算，人工智能交叉方向。在 SC, Journal of Chemical Theory and Computation, Journal of Computational Physics, Computer Physics Communications, Journal of Chemical Physics 等发表多篇期刊会议论文。作为核心人员，参与了包括 PWmat，LS3DF，PWDFT，PEXSI，DeePMD-kit 等多个科学计算软件的研发。