19th USENIX Symposium on Operating Systems Design and Implementation (OSDI '25)

July 7–9, 2025
Boston, MA, USA

Monday, July 7

Distributed Systems and Data Centers I

Basilisk: Using Provenance Invariants to Automate Proofs of Undecidable Protocols 1
Tony Nuda Zhang and Keshav Singh, *University of Michigan;* Tej Chajed, *University of Wisconsin-Madison;* Manos Kapritsos, *University of Michigan;* Bryan Parno, *Carnegie Mellon University*

Deriving Semantic Checkers from Tests to Detect Silent Failures in Production Distributed Systems 19
Chang Lou, *University of Virginia;* Dimas Shidqi Parikesit, *University of Virginia and Bandung Institute of Technology;* Yujin Huang, *The Pennsylvania State University;* Zhewen Yang and Senapati Diwangkara, *Johns Hopkins University;* Yuzhuo Jing, *University of Michigan;* Achmad Imam Kistijantoro, *Bandung Institute of Technology;* Ding Yuan, *University of Toronto;* Suman Nath, *Microsoft Research;* Peng Huang, *University of Michigan*

Picsou: Enabling Replicated State Machines to Communicate Efficiently 39
Reginald Frank, Micah Murray, Chawinphat Tankuranand, Junseo Yoo, Ethan Xu, and Natacha Crooks, *UC Berkeley;* Suyash Gupta, *University of Oregon;* Manos Kapritsos, *University of Michigan*

FineMem: Breaking the Allocation Overhead vs. Memory Waste Dilemma in Fine-Grained Disaggregated Memory Management 57
Xiaoyang Wang and Yongkun Li, *University of Science and Technology of China;* Kan Wu, *Google;* Wenzhe Zhu and Yuqi Li, *University of Science and Technology of China;* Yinlong Xu, *University of Science and Technology of China and Anhui Provincial Key Laboratory of High Performance Computing*

To PRI or Not To PRI, That's the question 75
Yun Wang, *Shanghai Jiao Tong University;* Liang Chen, Jie Ji, Xianting Tian, and Ben Luo, *Alibaba Group;* Zhixiang Wei, Zhibai Huang, and Kailiang Xu, *Shanghai Jiao Tong University;* Kaihuan Peng, Kaijie Guo, Ning Luo, Guangjian Wang, Shengdong Dai, Yibin Shen, and Jiesheng Wu, *Alibaba Group;* Zhengwei Qi, *Shanghai Jiao Tong University*

Enabling Efficient GPU Communication over Multiple NICs with FuseLink 91
Zhenghang Ren, Yuxuan Li, Zilong Wang, Xinyang Huang, Wenxue Li, Kaiqiang Xu, Xudong Liao, Yijun Sun, and Bowen Liu, *Hong Kong University of Science and Technology;* Han Tian, *University of Science and Technology of China;* Junxue Zhang, *Hong Kong University of Science and Technology;* Mingfei Wang, *MetaX Integrated Circuits;* Zhizhen Zhong, *Massachusetts Institute of Technology;* Guyue Liu, *Peking University;* Ying Zhang, *Meta;* Kai Chen, *Hong Kong University of Science and Technology*

Database Systems

Tigon: A Distributed Database for a CXL Pod 109
Yibo Huang, Haowei Chen, and Newton Ni, *The University of Texas at Austin;* Yan Sun, *University of Illinois Urbana–Champaign;* Vijay Chidambaram, Dixin Tang, and Emmett Witchel, *The University of Texas at Austin*

Mako: Speculative Distributed Transactions with Geo-Replication 129
Weihai Shen, *Stony Brook University;* Yang Cui, *Google;* Siddhartha Sen, *Microsoft Research;* Sebastian Angel, *University of Pennsylvania;* Shuai Mu, *Stony Brook University*

Quake: Adaptive Indexing for Vector Search 153
Jason Mohoney, Devesh Sarda, and Mengze Tang, *University of Wisconsin–Madison;* Shihabur Rahman Chowdhury and Anil Pacaci, *Apple;* Ihab F. Ilyas, *University of Waterloo;* Theodoros Rekatsinas, *Apple;* Shivaram Venkataraman, *University of Wisconsin–Madison*

Achieving Low-Latency Graph-Based Vector Search via Aligning Best-First Search Algorithm with SSD 171
Hao Guo and Youyou Lu, *Tsinghua University*

Skybridge: Bounded Staleness for Distributed Caches 187
Robert Lyerly, *Meta Platforms Inc.;* Scott Pruett, *unaffiliated;* Kevin Doherty and Greg Rogers, *Meta Platforms Inc.;* Nathan Bronson, *OpenAI;* John Hugg, *Meta Platforms Inc.*

AI + Systems I

KPerfIR: Towards a Open and Compiler-centric Ecosystem for GPU Kernel Performance Tooling on Modern AI Workloads 205
Yue Guan, *University of California, San Diego;* Yuanwei Fang, *Meta;* Keren Zhou, *George Mason University and OpenAI;* Corbin Robeck and Manman Ren, *Meta;* Zhongkai Yu, *University of California, San Diego;* Yufei Ding, *University of California, San Diego, and Meta;* Adnan Aziz, *Meta*

Mirage: A Multi-Level Superoptimizer for Tensor Programs 221
Mengdi Wu and Xinhao Cheng, *Carnegie Mellon University;* Shengyu Liu and Chunan Shi, *Peking University;* Jianan Ji and Man Kit Ao, *Carnegie Mellon University;* Praveen Velliengiri, *Pennsylvania State University;* Xupeng Miao, *Purdue University;* Oded Padon, *Weizmann Institute of Science;* Zhihao Jia, *Carnegie Mellon University*

QiMeng-Xpiler: Transcompiling Tensor Programs for Deep Learning Systems with a Neural-Symbolic Approach . . 239
Shouyang Dong, *University of Science and Technology of China, Cambricon Technologies, and Institute of Computing Technology, Chinese Academy of Sciences;* Yuanbo Wen, Jun Bi, Di Huang, and Jiaming Guo, *Institute of Computing Technology, Chinese Academy of Sciences;* Jianxing Xu and Ruibai Xu, *University of Science and Technology of China, Cambricon Technologies, and Institute of Computing Technology, Chinese Academy of Sciences;* Xinkai Song and Yifan Hao, *Institute of Computing Technology, Chinese Academy of Sciences;* Ling Li, *Institute of Software, Chinese Academy of Sciences, and University of Chinese Academy of Sciences;* Xuehai Zhou, *University of Science and Technology of China;* Tianshi Chen, *Cambricon Technologies;* Qi Guo, *Institute of Computing Technology, Chinese Academy of Sciences;* Yunji Chen, *Institute of Computing Technology, Chinese Academy of Sciences, and University of Chinese Academy of Sciences*

WaferLLM: Large Language Model Inference at Wafer Scale 257
Congjie He, Yeqi Huang, and Pei Mu, *University of Edinburgh;* Ziming Miao, Jilong Xue, Lingxiao Ma, and Fan Yang, *Microsoft Research;* Luo Mai, *University of Edinburgh*

Tuesday, July 8

AI + Systems II

BlitzScale: Fast and Live Large Model Autoscaling with *O*(1) Host Caching 275
Dingyan Zhang, Haotian Wang, Yang Liu, and Xingda Wei, *Shanghai Jiao Tong University;* Yizhou Shan, *Huawei Cloud;* Rong Chen and Haibo Chen, *Shanghai Jiao Tong University*

Bayesian Code Diffusion for Efficient Automatic Deep Learning Program Optimization. 295
Isu Jeong and Seulki Lee, *Ulsan National Institute of Science and Technology*

Training with Confidence: Catching Silent Errors in Deep Learning Training with Automated Proactive Checks. . . .313
Yuxuan Jiang, Ziming Zhou, Boyu Xu, Beijie Liu, Runhui Xu, and Peng Huang, *University of Michigan*

Neutrino: Fine-grained GPU Kernel Profiling via Programmable Probing 331
Songlin Huang and Chenshu Wu, *The University of Hong Kong*

Principles and Methodologies for Serial Performance Optimization 357
Sujin Park, Mingyu Guan, Xiang Cheng, and Taesoo Kim, *Georgia Institute of Technology*

Scheduling and Resource Management

Söze: One Network Telemetry Is All You Need for Per-flow Weighted Bandwidth Allocation at Scale 375
Weitao Wang and T. S. Eugene Ng, *Rice University*

Decouple and Decompose: Scaling Resource Allocation with DeDe 393
Zhiying Xu and Minlan Yu, *Harvard University;* Francis Y. Yan, *University of Illinois Urbana-Champaign*

Quantum Virtual Machines. 411
Runzhou Tao and Hongzheng Zhu, *University of Maryland, College Park;* Jason Nieh, *Columbia University;* Jianan Yao, *University of Toronto;* Ronghui Gu, *Columbia University*

QOS: Quantum Operating System 429
Emmanouil Giortamis, Francisco Romão, Nathaniel Tornow, and Pramod Bhatotia, *TU Munich*

Scalio: Scaling up DPU-based JBOF Key-value Store with NVMe-oF Target Offload 449
Xun Sun, Mingxing Zhang, Yingdi Shan, Kang Chen, and Jinlei Jiang, *Tsinghua University;* Yongwei Wu, *Tsinghua University, Quan Cheng Laboratory*

Distributed Systems and Data Centers II

Low End-to-End Latency atop a Speculative Shared Log with Fix-Ante Ordering 465
Shreesha G. Bhat, Tony Hong, Xuhao Luo, Jiyu Hu, Aishwarya Ganesan, and Ramnatthan Alagappan, *University of Illinois Urbana-Champaign*

Understanding Stragglers in Large Model Training Using What-if Analysis 483
Jinkun Lin, *New York University;* Ziheng Jiang, Zuquan Song, Sida Zhao, and Menghan Yu, *ByteDance Seed;* Zhanghan Wang, *New York University;* Chenyuan Wang, *ByteDance Seed;* Zuocheng Shi, *Zhejiang University;* Xiang Shi, *ByteDance;* Wei Jia, Zherui Liu, Shuguang Wang, Haibin Lin, and Xin Liu, *ByteDance Seed;* Aurojit Panda and Jinyang Li, *New York University*

Fork in the Road: Reflections and Optimizations for Cold Start Latency in Production Serverless Systems 499
Xiaohu Chai, *Tsinghua University; Ant Group;* Tianyu Zhou, *Ant Group;* Keyang Hu, *Tsinghua University;* Jianfeng Tan, Tiwei Bie, Anqi Shen, Dawei Shen, Qi Xing, Shun Song, Tongkai Yang, Le Gao, Feng Yu, and Zhengyu He, *Ant Group;* Dong Du and Yubin Xia, *Shanghai Jiao Tong University;* Kang Chen, *Tsinghua University;* Yu Chen, *Quan Cheng Laboratory; Tsinghua University*

Kamino: Efficient VM Allocation at Scale with Latency-Driven Cache-Aware Scheduling. 519
David Domingo, *Rutgers University;* Hugo Barbalho and Marco Molinaro, *Microsoft Research;* Kuan Liu, Abhisek Pan, David Dion, and Thomas Moscibroda, *Microsoft Azure;* Sudarsun Kannan, *Rutgers University;* Ishai Menache, *Microsoft Research*

Zen: Empowering Distributed Training with Sparsity-driven Data Synchronization 537
Zhuang Wang, *Rice University;* Zhaozhuo Xu, *Stevens Institute of Technology;* Jingyi Xi, *unaffiliated;* Yuke Wang, Anshumali Shrivastava, and T. S. Eugene Ng, *Rice University*

Kernel and Operating Systems I

Extending Applications Safely and Efficiently. 557
Yusheng Zheng, *UC Santa Cruz;* Tong Yu, *eunomia-bpf Community;* Yiwei Yang, *UC Santa Cruz;* Yanpeng Hu, *ShanghaiTech University;* Xiaozheng Lai, *South China University of Technology;* Dan Williams, *Virginia Tech;* Andi Quinn, *UC Santa Cruz*

Tintin: A Unified Hardware Performance Profiling Infrastructure to Uncover and Manage Uncertainty 575
Ao Li, Marion Sudvarg, Zihan Li, Sanjoy Baruah, Chris Gill, and Ning Zhang, *Washington University in St. Louis*

Building Bridges: Safe Interactions with Foreign Languages through Omniglot 595
Leon Schuermann and Jack Toubes, *Princeton University;* Tyler Potyondy and Pat Pannuto, *University of California San Diego;* Mae Milano and Amit Levy, *Princeton University*

KRR: Efficient and Scalable Kernel Record Replay. 615
Tianren Zhang, *SmartX;* Sishuai Gong and Pedro Fonseca, *Purdue University*

Deterministic Client: Enforcing Determinism on Untrusted Machine Code 633
Zachary Yedidia, *Stanford University;* Geoffrey Ramseyer, *Stanford University and Stellar Development Foundation;* David Mazieres, *Stanford University*

Wednesday, July 9

Kernel and Operating Systems II

Disentangling the Dual Role of NIC Receive Rings. 651
Boris Pismenny, *EPFL & NVIDIA;* Adam Morrison, *Tel Aviv University;* Dan Tsafrir, *Technion – Israel Institute of Technology*

XSched: Preemptive Scheduling for Diverse XPUs. 671
Weihang Shen, Mingcong Han, Jialong Liu, Rong Chen, and Haibo Chen, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University*

OS Rendering Service Made Parallel with Out-of-Order Execution and In-Order Commit 693
Yuanpei Wu and Dong Du, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education;* Chao Xu, *Fields Lab, Huawei Central Software Institute;* Yubin Xia and Yang Yu, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education;* Ming Fu, *Fields Lab, Huawei Central Software Institute;* Binyu Zang and Haibo Chen, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education*

EMT: An OS Framework for New Memory Translation Architectures **711**
Siyuan Chai, Jiyuan Zhang, Jongyul Kim, Alan Wang, Fan Chung, and Jovan Stojkovic, *University of Illinois Urbana-Champaign;* Weiwei Jia, *University of Rhode Island;* Dimitrios Skarlatos, *Carnegie Mellon University;* Josep Torrellas and Tianyin Xu, *University of Illinois Urbana-Champaign*

Tiered Memory Management Beyond Hotness **731**
Jinshu Liu, Hamid Hadian, Hanchen Xu, and Huaicheng Li, *Virginia Tech*

AI + Systems III

NanoFlow: Towards Optimal Large Language Model Serving Throughput **749**
Kan Zhu, *University of Washington;* Yufei Gao, *Tsinghua University and University of Washington;* Yilong Zhao, *University of Washington and University of California, Berkeley;* Liangyu Zhao, *University of Washington;* Gefei Zuo, *University of Michigan;* Yile Gu and Dedong Xie, *University of Washington;* Tian Tang and Qinyu Xu, *Tsinghua University and University of Washington;* Zihao Ye, Keisuke Kamahori, and Chien-Yu Lin, *University of Washington;* Ziren Wang, *Tsinghua University and University of Washington;* Stephanie Wang, Arvind Krishnamurthy, and Baris Kasikci, *University of Washington*

PipeThreader: Software-Defined Pipelining for Efficient DNN Execution **767**
Yu Cheng, Lei Wang, and Yining Shi, *School of Computer Science, Peking University;* Yuqing Xia, Lingxiao Ma, Jilong Xue, and Yang Wang, *Microsoft Research;* Zhiwen Mo, *Imperial College London and Microsoft Research;* Feiyang Chen, *Shanghai Jiao Tong University and Microsoft Research;* Fan Yang and Mao Yang, *Microsoft Research;* Zhi Yang, *School of Computer Science, Peking University*

WLB-LLM: Workload-Balanced 4D Parallelism for Large Language Model Training **785**
Zheng Wang, *University of California, San Diego, and Meta;* Anna Cai and Xinfeng Xie, *Meta;* Zaifeng Pan and Yue Guan, *University of California, San Diego;* Weiwei Chu, Jie Wang, Shikai Li, Jianyu Huang, Chris Cai, and Yuchen Hao, *Meta;* Yufei Ding, *University of California, San Diego, and Meta*

DecDEC: A Systems Approach to Advancing Low-Bit LLM Quantization **803**
Yeonhong Park, Jake Hyun, Hojoon Kim, and Jae W. Lee, *Seoul National University*

File and Storage Systems

Stripeless Data Placement for Erasure-Coded In-Memory Storage **821**
Jian Gao, Jiwu Shu, Bin Yan, and Yuhao Zhang, *Tsinghua University;* Keji Huang, *Huawei Technologies Co., Ltd*

PoWER Never Corrupts: Tool-Agnostic Verification of Crash Consistency and Corruption Detection **839**
Hayley LeBlanc, *University of Texas at Austin;* Jacob R. Lorch and Chris Hawblitzel, *Microsoft Research;* Cheng Huang and Yiheng Tao, *Microsoft;* Nickolai Zeldovich, *MIT CSAIL and Microsoft Research;* Vijay Chidambaram, *University of Texas at Austin*

Fast and Synchronous Crash Consistency with Metadata Write-Once File System **859**
Yanqi Pan, Wen Xia, Yifeng Zhang, Xiangyu Zou, and Hao Huang, *Harbin institute of Technology, Shenzhen;* Zhenhua Li, *Tsinghua University;* Chentao Wu, *Shanghai Jiao Tong University*

Decentralized, Epoch-based F2FS Journaling with Fine-grained Crash Recovery **879**
Yaotian Cui and Zhiqi Wang, *The Chinese University of Hong Kong, China;* Renhai Chen, *College of Intelligence and Computing, Tianjin University, China;* Zili Shao, *The Chinese University of Hong Kong, China*

***Okapi*: Decoupling Data Striping and Redundancy Grouping in Cluster File Systems** **897**
Sanjith Athlur and Timothy Kim, *Carnegie Mellon University;* Saurabh Kadekodi, *Google;* Francisco Maturana and Xavier Ramos, *Carnegie Mellon University;* Arif Merchant, *Google;* K. V. Rashmi and Gregory R. Ganger, *Carnegie Mellon University*

Privacy and Security

Compass: Encrypted Semantic Search with High Accuracy **915**
Jinhao Zhu, *UC Berkeley;* Liana Patel, *Stanford University;* Matei Zaharia and Raluca Ada Popa, *UC Berkeley*

Weave: Efficient and Expressive Oblivious Analytics at Scale **939**
Mahdi Soleimani, Grace Jia, and Anurag Khandelwal, *Yale University*

Paralegal: Practical Static Analysis for Privacy Bugs. **957**
Justus Adam, Carolyn Zech, Livia Zhu, Sreshtaa Rajesh, Nathan Harbison, Mithi Jethwa, Will Crichton, Shriram Krishnamurthi, and Malte Schwarzkopf, *Brown University*

MettEagle: Costs and Benefits of Implementing Containers on Microkernels **979**
Till Miemietz, Viktor Reusch, and Matthias Hille, *Barkhausen Institut;* Lars Wrenger, *Leibniz-Universität Hannover;* Jana Eisoldt, *Barkhausen Institut;* Jan Klötzke, *Kernkonzept GmbH;* Max Kurze, *Technische Universität Dresden;* Adam Lackorzynski, *Technische Universität Dresden and Kernkonzept GmbH;* Michael Roitzsch, *Barkhausen Institut;* Hermann Härtig, *Barkhausen Institut and Technische Universität Dresden*