18th USENIX Symposium on Operating Systems Design and Implementation (OSDI '24)

July 10–12, 2024
Santa Clara, CA, USA

Wednesday, July 10

Memory Management

Sabre: Hardware-Accelerated Snapshot Compression for Serverless MicroVMs 1
Nikita Lazarev and Varun Gohil, *MIT, CSAIL;* James Tsai, Andy Anderson, and Bhushan Chitlur, *Intel Labs;* Zhiru Zhang, *Cornell University;* Christina Delimitrou, *MIT, CSAIL*

Nomad: Non-Exclusive Memory Tiering via Transactional Page Migration 19
Lingfeng Xiang, Zhen Lin, Weishu Deng, Hui Lu, and Jia Rao, *The University of Texas at Arlington;* Yifan Yuan and Ren Wang, *Intel Labs*

Managing Memory Tiers with CXL in Virtualized Environments 37
Yuhong Zhong, *Columbia University, Microsoft Azure;* Daniel S. Berger, *Microsoft Azure, University of Washington;* Carl Waldspurger, *Carl Waldspurger Consulting;* Ryan Wee, *Columbia University;* Ishwar Agarwal, Rajat Agarwal, Frank Hady, and Karthik Kumar, *Intel;* Mark D. Hill, *University of Wisconsin–Madison;* Mosharaf Chowdhury, *University of Michigan;* Asaf Cidon, *Columbia University*

Harvesting Memory-bound CPU Stall Cycles in Software with MSH 57
Zhihong Luo, Sam Son, and Sylvia Ratnasamy, *UC Berkeley;* Scott Shenker, *UC Berkeley & ICSI*

A Tale of Two Paths: Toward a Hybrid Data Plane for Efficient Far-Memory Applications 77
Lei Chen, *University of Chinese Academy of Sciences;* Shi Liu, *UCLA;* Chenxi Wang, *University of Chinese Academy of Sciences;* Haoran Ma and Yifan Qiao, *UCLA;* Zhe Wang and Chenggang Wu, *University of Chinese Academy of Sciences;* Youyou Lu, *Tsinghua University;* Xiaobing Feng and Huimin Cui, *University of Chinese Academy of Sciences;* Shan Lu, *Microsoft Research;* Harry Xu, *UCLA*

DRust: Language-Guided Distributed Shared Memory with Fine Granularity, Full Transparency, and Ultra Efficiency 97
Haoran Ma, Yifan Qiao, Shi Liu, and Shan Yu, *UCLA;* Yuanjiang Ni, Qingda Lu, and Jiesheng Wu, *Alibaba Group;* Yiying Zhang, *UCSD;* Miryung Kim and Harry Xu, *UCLA*

Low-Latency LLM Serving

Taming Throughput-Latency Tradeoff in LLM Inference with *Sarathi-Serve* 117
Amey Agrawal, *Georgia Institute of Technology;* Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, and Bhargav Gulavani, *Microsoft Research India;* Alexey Tumanov, *Georgia Institute of Technology;* Ramachandran Ramjee, *Microsoft Research India*

ServerlessLLM: Low-Latency Serverless Inference for Large Language Models 135
Yao Fu, Leyang Xue, Yeqi Huang, and Andrei-Octavian Brabete, *University of Edinburgh;* Dmitrii Ustiugov, *NTU Singapore;* Yuvraj Patel and Luo Mai, *University of Edinburgh*

InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management 155
Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim, *Seoul National University*

Llumnix: Dynamic Scheduling for Large Language Model Serving 173
Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin, *Alibaba Group*

DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving 193
Yinmin Zhong and Shengyu Liu, *Peking University;* Junda Chen, *UC San Diego;* Jianbo Hu, *Peking University;* Yibo Zhu, *StepFun;* Xuanzhe Liu and Xin Jin, *Peking University;* Hao Zhang, *UC San Diego*

Distributed Systems

ACCL+: an FPGA-Based Collective Engine for Distributed Applications 211
Zhenhao He, Dario Korolija, Yu Zhu, and Benjamin Ramhorst, *Systems Group, ETH Zurich;* Tristan Laan, *University of Amsterdam;* Lucian Petrica and Michaela Blott, *AMD Research;* Gustavo Alonso, *Systems Group, ETH Zurich*

Beaver: Practical Partial Snapshots for Distributed Cloud Services 233
Liangcheng Yu, *University of Pennsylvania;* Xiao Zhang, *Shanghai Jiao Tong University;* Haoran Zhang, *University of Pennsylvania;* John Sonchack, *Princeton University;* Dan Ports, *Microsoft / University of Washington;* Vincent Liu, *University of Pennsylvania*

Fast and Scalable In-network Lock Management Using Lock Fission 251
Hanze Zhang, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Shanghai AI Laboratory; MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University;* Ke Cheng, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China;* Rong Chen, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Shanghai AI Laboratory; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China;* Haibo Chen, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China; Key Laboratory of System Software* (*Chinese Academy of Sciences*)

Chop Chop: Byzantine Atomic Broadcast to the Network Limit 269
Martina Camaioni, Rachid Guerraoui, Matteo Monti, Pierre-Louis Roman, Manuel Vidigueira, and Gauthier Voron, *EPFL*

Thursday, July 11

Deep Learning

Enabling Tensor Language Model to Assist in Generating High-Performance Tensor Programs for Deep Learning 289
Yi Zhai, *University of Science and Technology of China;* Sijia Yang, *Huawei Technologies Co., Ltd.;* Keyu Pan, *ByteDance Ltd.;* Renwei Zhang, *Huawei Technologies Co., Ltd.;* Shuo Liu, *University of Science and Technology of China;* Chao Liu and Zichun Ye, *Huawei Technologies Co., Ltd.;* Jianmin Ji, *University of Science and Technology of China;* Jie Zhao, *Hunan University;* Yu Zhang and Yanyong Zhang, *University of Science and Technology of China*

Ladder: Enabling Efficient Low-Precision Deep Learning Computing through Hardware-aware Tensor Transformation 307
Lei Wang, *University of Chinese Academy of Sciences & Microsoft Research;* Lingxiao Ma, Shijie Cao, Quanlu Zhang, and Jilong Xue, *Microsoft Research;* Yining Shi, *Peking University & Microsoft Research;* Ningxin Zheng, Ziming Miao, Fan Yang, Ting Cao, Yuqing Yang, and Mao Yang, *Microsoft Research*

Caravan: Practical Online Learning of In-Network ML Models with Labeling Agents 325
Qizheng Zhang, *Stanford University;* Ali Imran, *Purdue University;* Enkeleda Bardhi, *Sapienza University of Rome;* Tushar Swamy and Nathan Zhang, *Stanford University;* Muhammad Shahbaz, *Purdue University and University of Michigan;* Kunle Olukotun, *Stanford University*

nnScaler: Constraint-Guided Parallelization Plan Generation for Deep Learning Training 347
Zhiqi Lin, *University of Science and Technology of China;* Youshan Miao, Quanlu Zhang, Fan Yang, and Yi Zhu, *Microsoft Research;* Cheng Li, *University of Science and Technology of China;* Saeed Maleki, *xAI;* Xu Cao, Ning Shang, Yilei Yang, Weijiang Xu, and Mao Yang, *Microsoft Research;* Lintao Zhang, *BaseBit Technologies;* Lidong Zhou, *Microsoft Research*

ChameleonAPI: Automatic and Efficient Customization of Neural Networks for ML Applications 365
Yuhan Liu, *University of Chicago;* Chengcheng Wan, *East China Normal University;* Kuntai Du, Henry Hoffmann, and Junchen Jiang, *University of Chicago;* Shan Lu, *University of Chicago and Microsoft Research;* Michael Maire, *University of Chicago*

Operating Systems

SquirrelFS: using the Rust compiler to check file-system crash consistency 387
Hayley LeBlanc, Nathan Taylor, James Bornholt, and Vijay Chidambaram, *University of Texas at Austin*

High-throughput and Flexible Host Networking for Accelerated Computing 405
Athinagoras Skiadopoulos, Zhiqiang Xie, and Mark Zhao, *Stanford University;* Qizhe Cai and Saksham Agarwal, *Cornell University;* Jacob Adelmann, David Ahern, Carlo Contavalli, Michael Goldflam, Vitaly Mayatskikh, Raghu Raja, and Daniel Walton, *Enfabrica;* Rachit Agarwal, *Cornell University;* Shrijeet Mukherjee, *Enfabrica;* Christos Kozyrakis, *Stanford University*

IntOS: Persistent Embedded Operating System and Language Support for Multi-threaded Intermittent Computing 425
Yilun Wu, *Stony Brook University;* Byounguk Min, *Purdue University;* Mohannad Ismail and Wenjie Xiong, *Virginia Tech;* Changhee Jung, *Purdue University;* Dongyoon Lee, *Stony Brook University*

Data-flow Availability: Achieving Timing Assurance in Autonomous Systems 445
Ao Li and Ning Zhang, *Washington University in St. Louis*

Microkernel Goes General: Performance and Compatibility in the HongMeng Production Microkernel 465
Haibo Chen, *Huawei Central Software Institute and Shanghai Jiao Tong University;* Xie Miao, Ning Jia, Nan Wang, Yu Li, Nian Liu, Yutao Liu, Fei Wang, Qiang Huang, Kun Li, Hongyang Yang, Hui Wang, Jie Yin, Yu Peng, and Fengwei Xu, *Huawei Central Software Institute*

Cloud Computing

When will my ML Job finish? Toward providing Completion Time Estimates through Predictability-Centric Scheduling 487
Abdullah Bin Faisal, Noah Martin, Hafiz Mohsin Bashir, Swaminathan Lamelas, and Fahad R. Dogar, *Tufts University*

Optimizing Resource Allocation in Hyperscale Datacenters: Scalability, Usability, and Experiences 507
Neeraj Kumar, Pol Mauri Ruiz, Vijay Menon, Igor Kabiljo, Mayank Pundir, Andrew Newell, Daniel Lee, Liyuan Wang, and Chunqiang Tang, *Meta Platforms*

μSlope: High Compression and Fast Search on Semi-Structured Logs 529
Rui Wang, *YScope;* Devin Gibson, *YScope and University of Toronto;* Kirk Rodrigues, *YScope;* Yu Luo, *YScope, Uber, and University of Toronto;* Yun Zhang, Kaibo Wang, Yupeng Fu, and Ting Chen, *Uber;* Ding Yuan, *YScope and University of Toronto*

ServiceLab: Preventing Tiny Performance Regressions at Hyperscale through Pre-Production Testing 545
Mike Chow, *Meta Platforms;* Yang Wang, *Meta Platforms and The Ohio State University;* William Wang, Ayichew Hailu, Rohan Bopardikar, Bin Zhang, Jialiang Qu, David Meisner, Santosh Sonawane, Yunqi Zhang, Rodrigo Paim, Mack Ward, Ivor Huang, Matt McNally, Daniel Hodges, Zoltan Farkas, Caner Gocmen, Elvis Huang, and Chunqiang Tang, *Meta Platforms*

MAST: Global Scheduling of ML Training across Geo-Distributed Datacenters at Hyperscale 563
Arnab Choudhury, *Meta Platforms;* Yang Wang, *Meta Platforms and The Ohio State University;* Tuomas Pelkonen, *Meta Platforms;* Kutta Srinivasan, *LinkedIn;* Abha Jain, Shenghao Lin, Delia David, Siavash Soleimanifard, Michael Chen, Abhishek Yadav, Ritesh Tijoriwala, Denis Samoylov, and Chunqiang Tang, *Meta Platforms*

Formal Verification

Automatically Reasoning About How Systems Code Uses the CPU Cache 581
Rishabh Iyer, Katerina Argyraki, and George Candea, *EPFL*

VeriSMo: A Verified Security Module for Confidential VMs 599
Ziqiao Zhou, *Microsoft Research;* Anjali, *University of Wisconsin-Madison;* Weiteng Chen, *Microsoft Research;* Sishuai Gong, *Purdue University;* Chris Hawblitzel and Weidong Cui, *Microsoft Research*

Validating the eBPF Verifier via State Embedding 615
Hao Sun and Zhendong Su, *ETH Zurich*

Using Dynamically Layered Definite Releases for Verifying the RefFS File System 629
Mo Zou, Dong Du, and Mingkai Dong, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China;* Haibo Chen, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China; Huawei Technologies Co. Ltd*

Anvil: Verifying Liveness of Cluster Management Controllers **649**
Xudong Sun, Wenjie Ma, Jiawei Tyler Gu, and Zicheng Ma, *University of Illinois Urbana-Champaign;* Tej Chajed, *University of Wisconsin-Madison;* Jon Howell, Andrea Lattuada, and Oded Padon, *VMware Research;* Lalith Suresh, *Feldera;* Adriana Szekeres, *VMware Research;* Tianyin Xu, *University of Illinois Urbana-Champaign*

Friday, July 12

Cloud Security

DSig: Breaking the Barrier of Signatures in Data Centers **667**
Marcos K. Aguilera, *VMware Research Group;* Clément Burgelin, Rachid Guerraoui, and Antoine Murat, *École Polytechnique Fédérale de Lausanne (EPFL);* Athanasios Xygkis, *Oracle Labs;* Igor Zablotchi, *Mysten Labs*

Ransom Access Memories: Achieving Practical Ransomware Protection in Cloud with DeftPunk **687**
Zhongyu Wang, Yaheng Song, Erci Xu, Haonan Wu, Guangxun Tong, Shizhuo Sun, Haoran Li, Jincheng Liu, Lijun Ding, Rong Liu, Jiaji Zhu, and Jiesheng Wu, *Alibaba Group*

Secret Key Recovery in a Global-Scale End-to-End Encryption System **703**
Graeme Connell, *Signal Messenger;* Vivian Fang, *UC Berkeley;* Rolfe Schmidt, *Signal Messenger;* Emma Dauterman and Raluca Ada Popa, *UC Berkeley*

***Flock*: A Framework for *Deploying* On-Demand Distributed Trust** **721**
Darya Kaviani and Sijun Tan, *UC Berkeley;* Pravein Govindan Kannan, *IBM Research;* Raluca Ada Popa, *UC Berkeley*

Data Management

FairyWren: A Sustainable Cache for Emerging Write-Read-Erase Flash Interfaces **745**
Sara McAllister and Yucong "Sherry" Wang, *Carnegie Mellon University;* Benjamin Berg, *UNC Chapel Hill;* Daniel S. Berger, *Microsoft Azure and University of Washington;* George Amvrosiadis, Nathan Beckmann, and Gregory R. Ganger, *Carnegie Mellon University*

Massively Parallel Multi-Versioned Transaction Processing **765**
Shujian Qian and Ashvin Goel, *University of Toronto*

Burstable Cloud Block Storage with Data Processing Units **783**
Junyi Shu, *School of Computer Science, Peking University and Alibaba Cloud;* Kun Qian and Ennan Zhai, *Alibaba Cloud;* Xuanzhe Liu and Xin Jin, *School of Computer Science, Peking University*

Motor: Enabling Multi-Versioning for Distributed Transactions on Disaggregated Memory **801**
Ming Zhang, Yu Hua, and Zhijun Yang, *Wuhan National Laboratory for Optoelectronics, School of Computer, Huazhong University of Science and Technology*

Analysis of Correctness

Detecting Logic Bugs in Database Engines via Equivalent Expression Transformation **821**
Zu-Ming Jiang and Zhendong Su, *ETH Zurich*

Inductive Invariants That Spark Joy: Using Invariant Taxonomies to Streamline Distributed Protocol Proofs. **837**
Tony Nuda Zhang, *University of Michigan;* Travis Hance, *Carnegie Mellon University;* Manos Kapritsos, *University of Michigan;* Tej Chajed, *University of Wisconsin–Madison;* Bryan Parno, *Carnegie Mellon University*

Performance Interfaces for Hardware Accelerators **855**
Jiacheng Ma, Rishabh Iyer, Sahand Kashani, Mahyar Emami, Thomas Bourgeat, and George Candea, *EPFL*

IronSpec: Increasing the Reliability of Formal Specifications **875**
Eli Goldweber, Weixin Yu, Seyed Armin Vakil Ghahani, and Manos Kapritsos, *University of Michigan*

Identifying On-/Off-CPU Bottlenecks Together with Blocked Samples **893**
Minwoo Ahn and Jeongmin Han, *Sungkyunkwan University;* Youngjin Kwon, *Korea Advanced Institute of Science and Technology (KAIST);* Jinkyu Jeong, *Yonsei University*

ML Scheduling

dLoRA: Dynamically Orchestrating Requests and Adapters for LoRA LLM Serving 911
Bingyang Wu, Ruidong Zhu, and Zili Zhang, *School of Computer Science, Peking University;* Peng Sun, *Shanghai AI Lab;* Xuanzhe Liu and Xin Jin, *School of Computer Science, Peking University*

Parrot: Efficient Serving of LLM-based Applications with Semantic Variable. 929
Chaofan Lin, *Shanghai Jiao Tong University;* Zhenhua Han, Chengruidong Zhang, Yuqing Yang, and Fan Yang, *Microsoft Research;* Chen Chen, *Shanghai Jiao Tong University;* Lili Qiu, *Microsoft Research*

Usher: Holistic Interference Avoidance for Resource Optimized ML Inference 947
Sudipta Saha Shubha and Haiying Shen, *University of Virginia;* Anand Iyer, *Georgia Institute of Technology*

Fairness in Serving Large Language Models. 965
Ying Sheng, *UC Berkeley and Stanford University;* Shiyi Cao, Dacheng Li, Banghua Zhu, and Zhuohan Li, *UC Berkeley;* Danyang Zhuo, *Duke University;* Joseph E. Gonzalez and Ion Stoica, *UC Berkeley*

MonoNN: Enabling a New Monolithic Optimization Space for Neural Network Inference Tasks on Modern GPU-Centric Architectures 989
Donglin Zhuang, *The University of Sydney;* Zhen Zheng, *Alibaba Group;* Haojun Xia, *The University of Sydney;* Xiafei Qiu, Junjie Bai, and Wei Lin, *Alibaba Group;* Shuaiwen Leon Song, *The University of Sydney*