

# 2023 USENIX Annual Technical Conference

July 10–12, 2023

Boston, MA, USA

## Monday, July 10

### Security and Privacy

**Bifrost: Analysis and Optimization of Network I/O Tax in Confidential Virtual Machines** ..... 1

Dingji Li, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China; MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University*; Zeyu Mi, Chenhui Ji, Yifan Tan, and Binyu Zang, *Institute of Parallel and Distributed Systems, SEIEE, Shanghai Jiao Tong University; Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China*; Haibing Guan, *Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University*; 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*

**SecretFlow-SPU: A Performant and User-Friendly Framework for Privacy-Preserving Machine Learning** .....17

Junming Ma, Yancheng Zheng, Jun Feng, Derun Zhao, Haoqi Wu, Wenjing Fang, Jin Tan, Chaofan Yu, Benyu Zhang, and Lei Wang, *Ant Group*

**Portunus: Re-imagining Access Control in Distributed Systems**..... 35

Watson Ladd, *Akamai*; Tanya Verma, *Cloudflare*; Marloes Venema, *University of Wuppertal*; Armando Faz-Hernández, *Cloudflare*; Brendan McMillion; Avani Wildani and Nick Sullivan, *Cloudflare*

### Searching Graphs

**GLogS: Interactive Graph Pattern Matching Query At Large Scale** ..... 53

Longbin Lai, *Alibaba Group, China*; Yufan Yang, *The Chinese University of Hong Kong, Shenzhen*; Zhibin Wang, *Nanjing University*; Yuxuan Liu and Haotian Ma, *The Chinese University of Hong Kong, Shenzhen*; Sijie Shen, Bingqing Lyu, Xiaoli Zhou, Wenyuan Yu, and Zhengping Qian, *Alibaba Group, China*; Chen Tian and Sheng Zhong, *Nanjing University*; Yeh-Ching Chung, *The Chinese University of Hong Kong, Shenzhen*; Jingren Zhou, *Alibaba Group, China*

**Cyclosa: Redundancy-Free Graph Pattern Mining via Set Dataflow** ..... 71

Chuangyi Gui, *National Engineering Research Center for Big Data Technology and System/Service Computing Technology and System Lab/Cluster and Grid Computing Lab, Huazhong University of Science and Technology, China; Zhejiang Lab, China*; Xiaofei Liao, *National Engineering Research Center for Big Data Technology and System/Service Computing Technology and System Lab/Cluster and Grid Computing Lab, Huazhong University of Science and Technology, China*; Long Zheng, *National Engineering Research Center for Big Data Technology and System/Service Computing Technology and System Lab/Cluster and Grid Computing Lab, Huazhong University of Science and Technology, China; Zhejiang Lab, China*; Hai Jin, *National Engineering Research Center for Big Data Technology and System/Service Computing Technology and System Lab/Cluster and Grid Computing Lab, Huazhong University of Science and Technology, China*

**SOWalker: An I/O-Optimized Out-of-Core Graph Processing System for Second-Order Random Walks**..... 87

Yutong Wu, Zhan Shi, Shicai Huang, Zhipeng Tian, Pengwei Zuo, Peng Fang, Fang Wang, and Dan Feng, *Wuhan National Laboratory for Optoelectronics Huazhong University of Science and Technology*

### Deduplication

**Light-Dedup: A Light-weight Inline Deduplication Framework for Non-Volatile Memory File Systems** .....101

Jiansheng Qiu, Yanqi Pan, Wen Xia, Xiaojia Huang, Wenjun Wu, Xiangyu Zou, and Shiyi Li, *Harbin Institute of Technology, Shenzhen*; Yu Hua, *Huazhong University of Science and Technology*

**TiDedup: A New Distributed Deduplication Architecture for Ceph** .....117

Myoungwon Oh and Sungmin Lee, *Samsung Electronics Co.*; Samuel Just, *IBM*; Young Jin Yu and Duck-Ho Bae, *Samsung Electronics Co.*; Sage Weil, *Ceph Foundation*; Sangyeun Cho, *Samsung Electronics Co.*; Heon Y. Yeom, *Seoul National University*

**LoopDelta: Embedding Locality-aware Opportunistic Delta Compression in Inline Deduplication for Highly Efficient Data Reduction** ..... 133  
Yucheng Zhang, *School of Mathematics and Computer Sciences, Nanchang University and Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology*; Hong Jiang, *Department of Computer Science and Engineering, University of Texas at Arlington*; Dan Feng, *Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology*; Nan Jiang, *School of Information Engineering, East China Jiaotong University*; Taorong Qiu and Wei Huang, *School of Mathematics and Computer Sciences, Nanchang University*

## Structuring Graphs

**TC-GNN: Bridging Sparse GNN Computation and Dense Tensor Cores on GPUs** .....149  
Yuke Wang, Boyuan Feng, Zheng Wang, Guyue Huang, and Yufei Ding, *University of California, Santa Barbara*

**Legion: Automatically Pushing the Envelope of Multi-GPU System for Billion-Scale GNN Training** ..... 165  
Jie Sun, *Collaborative Innovation Center of Artificial Intelligence, Zhejiang University, China*; Li Su, *Alibaba Group*; Zuo Cheng Shi, *Collaborative Innovation Center of Artificial Intelligence, Zhejiang University, China*; Wenting Shen, *Alibaba Group*; Zeke Wang, *Collaborative Innovation Center of Artificial Intelligence, Zhejiang University, China*; Lei Wang, *Alibaba Group*; Jie Zhang, *Collaborative Innovation Center of Artificial Intelligence, Zhejiang University, China*; Yong Li, Wenyuan Yu, and Jingren Zhou, *Alibaba Group*; Fei Wu, *Collaborative Innovation Center of Artificial Intelligence, Zhejiang University, China and Shanghai Institute for Advanced Study of Zhejiang University, China*

**Bridging the Gap between Relational OLTP and Graph-based OLAP** ..... 181  
Sijie Shen, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University and Alibaba Group*; Zihang Yao and Lin Shi, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University*; Lei Wang, Longbin Lai, Qian Tao, and Li Su, *Alibaba Group*; Rong Chen, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University and Shanghai AI Laboratory*; Wenyuan Yu, *Alibaba Group*; Haibo Chen and Binyu Zang, *Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University*; Jingren Zhou, *Alibaba Group*

## Placement and Fault Tolerance

**Comosum: An Extensible, Reconfigurable, and Fault-Tolerant IoT Platform for Digital Agriculture** ..... 197  
Gloire Rubambiza, Shiang-Wan Chin, Mueed Rehman, Sachille Atapattu, José F. Martínez, and Hakim Weatherspoon, *Cornell University*

**oakestra: A Lightweight Hierarchical Orchestration Framework for Edge Computing** ..... 215  
Giovanni Bartolomeo, Mehdi Yosofie, Simon Baurle, Oliver Haluszczynski, Nitinder Mohan, and Jörg Ott, *Technical University of Munich, Germany*

**Explore Data Placement Algorithm for Balanced Recovery Load Distribution.** ..... 233  
Yingdi Shan, *Zhongguancun Laboratory and Tsinghua University*; Kang Chen and Yongwei Wu, *Tsinghua University*

## Updating Code

**LUCI: Loader-based Dynamic Software Updates for Off-the-shelf Shared Objects.** ..... 241  
Bernhard Heinloth, Peter Wagemann, and Wolfgang Schröder-Preikschat, *Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany*

**MELF: Multivariant Executables for a Heterogeneous World** ..... 257  
Dominik Töllner, *Leibniz Universität Hannover*; Christian Dietrich, *Hamburg University of Technology*; Illia Ostapysyn, Florian Rommel, and Daniel Lohmann, *Leibniz Universität Hannover*

**APRON: Authenticated and Progressive System Image Renovation** ..... 275  
Sangho Lee, *Microsoft Research*

**zpoline: a system call hook mechanism based on binary rewriting.** ..... 293  
Kenichi Yasukata, Hajime Tazaki, and Pierre-Louis Aublin, *IIJ Research Laboratory*; Kenta Ishiguro, *Hosei University*

## Tuesday, July 11

### Serverless

**Sponge: Fast Reactive Scaling for Stream Processing with Serverless Frameworks** . . . . . 301  
Won Wook Song, *Seoul National University*; Taegeon Um, *Samsung Research*; Sameh Elnikety, *Microsoft Research*;  
Myeongjae Jeon, *UNIST*; Byung-Gon Chun, *Seoul National University and FriendliAI*

**On-demand Container Loading in AWS Lambda** . . . . . 315  
Marc Brooker, Mike Danilov, Chris Greenwood, and Phil Piwonka, *Amazon Web Services*

**Decentralized and Stateful Serverless Computing on the Internet Computer Blockchain** . . . . . 329  
Maksym Arutyunyan, Andriy Berestovskyy, Adam Bratschi-Kaye, Ulan Degenbaev, Manu Drijvers, Islam El-Ashi,  
Stefan Kaestle, Roman Kashitsyn, Maciej Kot, Yvonne-Anne Pignolet, Rostislav Rumenov, Dimitris Sarlis, Alin Sinpalean,  
Alexandru Uta, Bogdan Warinschi, and Alexandra Zapuc, *DFINITY, Zurich*

### Troubleshooting and Measurement

**PINOLO: Detecting Logical Bugs in Database Management Systems with Approximate Query Synthesis** . . . . . 345  
Zongyin Hao and Quanfeng Huang, *School of Informatics, Xiamen University*; Chengpeng Wang, *The Hong Kong University  
of Science and Technology*; Jianfeng Wang, *University of Southern California*; Yushan Zhang, *Tencent Inc.*; Rongxin Wu,  
*School of Informatics, Xiamen University*; Charles Zhang, *The Hong Kong University of Science and Technology*

**AutoARTS: Taxonomy, Insights and Tools for Root Cause Labelling of Incidents in Microsoft Azure** . . . . . 359  
Pradeep Dogga, *UCLA*; Chetan Bansal, Richard Costleigh, Gopinath Jayagopal, Suman Nath, and Xuchao Zhang, *Microsoft*

**Avoiding the Ordering Trap in Systems Performance Measurement** . . . . . 373  
Dmitry Duplyakin and Nikhil Ramesh, *University of Utah*; Carina Imburgia, *University of Washington*; Hamza Fathallah  
Al Sheikh, Semil Jain, Prikshit Tekta, Aleksander Maricq, Gary Wong, and Robert Ricci, *University of Utah*

### Cloud and Microservices

**AWARE: Automate Workload Autoscaling with Reinforcement Learning in Production Cloud Systems** . . . . . 387  
Haoran Qiu and Weichao Mao, *University of Illinois at Urbana-Champaign*; Chen Wang, Hubertus Franke, and Alaa Youssef,  
*IBM Research*; Zbigniew T. Kalbarczyk, Tamer Başar, and Ravishankar K. Iyer, *University of Illinois at Urbana-Champaign*

**Nodens: Enabling Resource Efficient and Fast QoS Recovery of Dynamic Microservice Applications in Datacenters** . . 403  
Jiuchen Shi, Hang Zhang, Zhixin Tong, Quan Chen, Kaihua Fu, and Minyi Guo, *Department of Computer Science and  
Engineering, Shanghai Jiao Tong University*

**Lifting the veil on Meta’s microservice architecture: Analyses of topology and request workflows** . . . . . 419  
Darby Huye, *Tufts University, Meta*; Yuri Shkuro, *Meta*; Raja R. Sambasivan, *Tufts University*

### Distributed Storage

**Tectonic-Shift: A Composite Storage Fabric for Large-Scale ML Training** . . . . . 433  
Mark Zhao, *Stanford University and Meta*; Satadru Pan, Niket Agarwal, Zhaoduo Wen, David Xu, Anand Natarajan, Pavan  
Kumar, Shiva Shankar P, Ritesh Tijoriwala, Karan Asher, Hao Wu, Aarti Basant, Daniel Ford, Delia David, Nezhil Yigitbasi,  
Pratap Singh, and Carole-Jean Wu, *Meta*; Christos Kozyrakis, *Stanford University*

**Calcspar: A Contract-Aware LSM Store for Cloud Storage with Low Latency Spikes** . . . . . 451  
Yuanhui Zhou and Jian Zhou, *WNLO, Huazhong University of Science and Technology, Wuhan, Hubei, China*; Shuning  
Chen, *PingCAP, China*; Peng Xu, *Research Center for Graph Computing, Zhejiang Lab, Hangzhou, Zhejiang, China*;  
Peng Wu, *WNLO, Huazhong University of Science and Technology, Wuhan, Hubei, China*; Yanguang Wang and Xian  
Liu, *PingCAP, China*; Ling Zhan, *Division of Information Science and Technology, Wenhua University, Wuhan, China*;  
Jiguang Wan, *WNLO, Huazhong University of Science and Technology, Wuhan, Hubei, China*

**Adaptive Online Cache Capacity Optimization via Lightweight Working Set Size Estimation at Scale** . . . . . 467  
Rong Gu, Simian Li, Haipeng Dai, Hancheng Wang, and Yili Luo, *State Key Laboratory for Novel Software Technology,  
Nanjing University, Nanjing 210023, China*; Bin Fan, *Alluxio Inc*; Ran Ben Basat, *University College London*; Ke Wang,  
*Meta Inc*; Zhenyu Song, *Princeton University*; Shouwei Chen and Beinan Wang, *Alluxio Inc*; Yihua Huang and Guihai Chen,  
*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China*

## Hardware and Software for Security and Performance

**SAGE: Software-based Attestation for GPU Execution** ..... 485  
Andrei Ivanov and Benjamin Rothenberger, *ETH Zürich*; Arnaud Dethise and Marco Canini, *KAUST*; Torsten Hoefler and Adrian Perrig, *ETH Zürich*

**Confidential Computing within an AI Accelerator** ..... 501  
Kapil Vaswani, Stavros Volos, Cédric Fournet, Antonio Nino Diaz, and Ken Gordon, *Microsoft*; Balaji Vembu, *Meta*; Sam Webster and David Chisnall, *Microsoft*; Saurabh Kulkarni, *Lucata Systems*; Graham Cunningham, *XTX Markets*; Richard Osborne, *Graphcore*; Daniel Wilkinson, *Imagination Technologies*

**Arbitor: A Numerically Accurate Hardware Emulation Tool for DNN Accelerators** ..... 519  
Chenhao Jiang and Anand Jayarajan, *University of Toronto and Vector Institute*; Hao Lu, *University of Toronto*; Gennady Pekhimenko, *University of Toronto and Vector Institute*

## Networking

**oBBR: Optimize Retransmissions of BBR Flows on the Internet** ..... 537  
Pengqiang Bi, Mengbai Xiao, Dongxiao Yu, and Guanghui Zhang, *Shandong University*; Jian Tong, Jingchao Liu, and Yijun Li, *BaishanCloud*

**Bridging the Gap between QoE and QoS in Congestion Control: A Large-scale Mobile Web Service Perspective** .... 553  
Jia Zhang, *Tsinghua University, Zhongguancun Laboratory, Beijing National Research Center for Information Science and Technology*; Yixuan Zhang, *Tsinghua University, Beijing National Research Center for Information Science and Technology*; Enhuan Dong, *Tsinghua University, Quan Cheng Laboratory, Beijing National Research Center for Information Science and Technology*; Yan Zhang, Shaorui Ren, and Zili Meng, *Tsinghua University, Beijing National Research Center for Information Science and Technology*; Mingwei Xu, *Tsinghua University, Quan Cheng Laboratory, Beijing National Research Center for Information Science and Technology*; Xiaotian Li, Zongzhi Hou, and Zhicheng Yang, *Meituan Inc.*; Xiaoming Fu, *University of Goettingen*

**FarReach: Write-back Caching in Programmable Switches** ..... 571  
Siyuan Sheng and Huancheng Puyang, *The Chinese University of Hong Kong*; Qun Huang, *Peking University*; Lu Tang, *Xiamen University*; Patrick P. C. Lee, *The Chinese University of Hong Kong*

## Memory-Related Hardware and Software

**CXL-ANNS: Software-Hardware Collaborative Memory Disaggregation and Computation for Billion-Scale Approximate Nearest Neighbor Search** ..... 585  
Junhyeok Jang, *Computer Architecture and Memory Systems Laboratory, KAIST*; Hanjin Choi, *Computer Architecture and Memory Systems Laboratory, KAIST and Panmnnesia, Inc.*; Hanyeoreum Bae and Seungjun Lee, *Computer Architecture and Memory Systems Laboratory, KAIST*; Miryeong Kwon and Myoungsoo Jung, *Computer Architecture and Memory Systems Laboratory, KAIST and Panmnnesia, Inc.*

**Overcoming the Memory Wall with CXL-Enabled SSDs**..... 601  
Shao-Peng Yang, *Syracuse University*; Minjae Kim, *DGIST*; Sanghyun Nam, *Soongsil University*; Juhung Park, *DGIST*; Jin-yong Choi and Eeye Hyun Nam, *FADU Inc.*; Eunji Lee, *Soongsil University*; Sungjin Lee, *DGIST*; Bryan S. Kim, *Syracuse University*

**STRYX: Exploiting SmartNIC Capability to Reduce Datacenter Memory Tax**..... 619  
Houxiang Ji, *University of Illinois Urbana-Champaign*; Mark Mansi, *University of Wisconsin-Madison*; Yan Sun, *University of Illinois Urbana-Champaign*; Yifan Yuan, *Intel Labs*; Jinghan Huang and Reese Kuper, *University of Illinois Urbana-Champaign*; Michael M. Swift, *University of Wisconsin-Madison*; Nam Sung Kim, *University of Illinois Urbana-Champaign*

## Deployed Networking

**Change Management in Physical Network Lifecycle Automation** ..... 635  
Mohammad Al-Fares, Virginia Beauregard, Kevin Grant, Angus Griffith, Jahangir Hasan, Chen Huang, Quan Leng, Jiayao Li, and Alexander Lin, *Google*; Zhuotao Liu, *Tsinghua University*; Ahmed Mansy, *Google*; Bill Martinusen, *Formerly at Google*; Nikil Mehta, Jeffrey C. Mogul, Andrew Narver, and Anshul Nigam, *Google*; Melanie Obenberger, *Formerly at Google*; Sean Smith, *Databricks*; Kurt Steinkraus, Sheng Sun, Edward Thiele, and Amin Vahdat, *Google*



**AAsclepius: Monitoring, Diagnosing, and Detouring at the Internet Peering Edge** ..... 655  
Kaicheng Yang and Yuanpeng Li, *National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University and Peng Cheng Laboratory, Shenzhen, China*; Sheng Long, *National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University and Huawei Cloud Computing Technologies Co., Ltd., China*; Tong Yang, *National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University and Peng Cheng Laboratory, Shenzhen, China*; Ruijie Miao and Yikai Zhao, *National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University*; Chaoyang Ji, Penghui Mi, Guodong Yang, Qiong Xie, Hao Wang, Yinhua Wang, Bo Deng, Zhiqiang Liao, Chengqiang Huang, Yongqiang Yang, Xiang Huang, Wei Sun, and Xiaoping Zhu, *Huawei Cloud Computing Technologies Co., Ltd., China*

**Deploying User-space TCP at Cloud Scale with LUNA** ..... 673  
Lingjun Zhu, Yifan Shen, Erci Xu, Bo Shi, Ting Fu, Shu Ma, Shuguang Chen, Zhongyu Wang, Haonan Wu, Xingyu Liao, Zhendan Yang, Zhongqing Chen, Wei Lin, Yijun Hou, Rong Liu, Chao Shi, Jiaji Zhu, and Jiessheng Wu, *Alibaba Group*

## Key-Value Stores

**RubbleDB: CPU-Efficient Replication with NVMe-oF** ..... 689  
Haoyu Li, Sheng Jiang, and Chen Chen, *Columbia University*; Ashwini Raina, *Princeton University*; Xingyu Zhu, Changxu Luo, and Asaf Cidon, *Columbia University*

**Distributed Transactions at Scale in Amazon DynamoDB** ..... 705  
Joseph Idziorek, Alex Keyes, Colin Lazier, Somu Perianayagam, Prithvi Ramanathan, James Christopher Sorenson III, Doug Terry, and Akshat Vig, *Amazon Web Services*

## Security: Attacks

**Prefix Siphoning: Exploiting LSM-Tree Range Filters For Information Disclosure** ..... 719  
Adi Kaufman, *Tel Aviv University*; Moshik Hershcovitch, *Tel Aviv University & IBM Research*; Adam Morrison, *Tel Aviv University*

**EPF: Evil Packet Filter** ..... 735  
Di Jin, Vaggelis Atlidakis, and Vasileios P. Kemerlis, *Brown University*

## Wednesday, July 12

### Virtual Machines

**Translation Pass-Through for Near-Native Paging Performance in VMs** ..... 753  
Shai Bergman and Mark Silberstein, *Technion*; Takahiro Shinagawa, *University of Tokyo*; Peter Pietzuch and Lluís Vilanova, *Imperial College London*

**Efficient Memory Overcommitment for I/O Passthrough Enabled VMs via Fine-grained Page Meta-data Management** ..... 769  
Yaohui Wang, Ben Luo, and Yibin Shen, *Alibaba Group*

**LPNS: Scalable and Latency-Predictable Local Storage Virtualization for Unpredictable NVMe SSDs in Clouds** ... 785  
Bo Peng, Cheng Guo, Jianguo Yao, and Haibing Guan, *Shanghai Jiao Tong University*

### Persistent Memory

**P<sup>2</sup>CACHE: Exploring Tiered Memory for In-Kernel File Systems Caching** ..... 801  
Zhen Lin, *Binghamton University*; Lingfeng Xiang and Jia Rao, *The University of Texas at Arlington*; Hui Lu, *Binghamton University*

**Revisiting Secondary Indexing in LSM-based Storage Systems with Persistent Memory** ..... 817  
Jing Wang, Youyou Lu, Qing Wang, Yuhao Zhang, and Jiwu Shu, *Department of Computer Science and Technology, Tsinghua University and Beijing National Research Center for Information Science and Technology (BNRist)*

**Zhuque: Failure is Not an Option, it's an Exception** ..... 833  
George Hodgkins, *University of Colorado, Boulder*; Yi Xu and Steven Swanson, *University of California, San Diego*; Joseph Izraelevitz, *University of Colorado, Boulder*

## Offloading and Scheduling

**ENVPIPE: Performance-preserving DNN Training Framework for Saving Energy** ..... 851  
Sangjin Choi and Inho Koo, *KAIST*; Jeongseob Ahn, *Ajou University*; Myeongjae Jeon, *UNIST*; Youngjin Kwon, *KAIST*

**Decentralized Application-Level Adaptive Scheduling for Multi-Instance DNNs on Open Mobile Devices**..... 865  
Hsin-Hsuan Sung and Jou-An Chen, *Department of Computer Science, North Carolina State University*; Wei Niu, Jiexiong Guan, and Bin Ren, *Department of Computer Science, William & Mary*; Xipeng Shen, *Department of Computer Science, North Carolina State University*

**UnFaaSener: Latency and Cost Aware Offloading of Functions from Serverless Platforms**..... 879  
Ghazal Sadeghian and Mohamed Elsakhawy, *University of British Columbia*; Mohanna Shahrads, *McGill University*; Joe Hattori, *University of Tokyo*; Mohammad Shahrads, *University of British Columbia*

## Kernel and Concurrency

**LLFREE: Scalable and Optionally-Persistent Page-Frame Allocation**..... 897  
Lars Wrenger, Florian Rommel, and Alexander Halbuer, *Leibniz Universität Hannover*; Christian Dietrich, *Hamburg University of Technology*; Daniel Lohmann, *Leibniz Universität Hannover*

**SINGULARFS: A Billion-Scale Distributed File System Using a Single Metadata Server** ..... 915  
Hao Guo, Youyou Lu, Wenhao Lv, Xiaojian Liao, Shaoxun Zeng, and Jiwu Shu, *Tsinghua University*

**The Hitchhiker's Guide to Operating Systems** ..... 929  
Yanyan Jiang, *Nanjing University*

## Optimizing ML

**Accelerating Distributed MoE Training and Inference with Lina** ..... 945  
Jiamin Li, *City University of Hong Kong*; Yimin Jiang, *ByteDance Inc.*; Yibo Zhu, *Unaffiliated*; Cong Wang, *City University of Hong Kong*; Hong Xu, *The Chinese University of Hong Kong*

**SMARTMOE: Efficiently Training Sparsely-Activated Models through Combining Offline and Online Parallelization** ..961  
Mingshu Zhai, Jiaao He, Zixuan Ma, Zan Zong, Runqing Zhang, and Jidong Zhai, *Tsinghua University*

**MSRL: Distributed Reinforcement Learning with Dataflow Fragments**..... 977  
Huanzhou Zhu, *Imperial College London*; Bo Zhao, *Imperial College London and Aalto University*; Gang Chen, Weifeng Chen, Yijie Chen, and Liang Shi, *Huawei Technologies Co., Ltd.*; Yaodong Yang, *Peking University*; Peter Pietzuch, *Imperial College London*; Lei Chen, *Hong Kong University of Science and Technology*

## GPU

**Beware of Fragmentation: Scheduling GPU-Sharing Workloads with Fragmentation Gradient Descent** ..... 995  
Qizhen Weng and Lingyun Yang, *Hong Kong University of Science and Technology*; Yinghao Yu, *Alibaba Group and Hong Kong University of Science and Technology*; Wei Wang, *Hong Kong University of Science and Technology*; Xiaochuan Tang, Guodong Yang, and Liping Zhang, *Alibaba Group*

**Towards Iterative Relational Algebra on the GPU** ..... 1009  
Ahmedur Rahman Shovon and Thomas Gilray, *University of Alabama at Birmingham*; Kristopher Micinski, *Syracuse University*; Sidharth Kumar, *University of Alabama at Birmingham*

**VectorVisor: A Binary Translation Scheme for Throughput-Oriented GPU Acceleration** .....1017  
Samuel Ginzburg, *Princeton University*; Mohammad Shahrads, *University of British Columbia*; Michael J. Freedman, *Princeton University*