

OSDI '20: 14th USENIX Symposium on Operating Systems Design and Implementation

November 4–6, 2020

Wednesday, November 4

Correctness

- Theseus: an Experiment in Operating System Structure and State Management** 1
Kevin Boos, *Rice University*; Namitha Liyanage, *Yale University*; Ramla Ijaz, *Rice University*; Lin Zhong, *Yale University*
- RedLeaf: Isolation and Communication in a Safe Operating System** 21
Vikram Narayanan, Tianjiao Huang, David Detweiler, Dan Appel, and Zhaofeng Li, *University of California, Irvine*; Gerd Zellweger, *VMware Research*; Anton Burtsev, *University of California, Irvine*
- Specification and verification in the field: Applying formal methods to BPF just-in-time compilers in the Linux kernel** 41
Luke Nelson, Jacob Van Geffen, Emina Torlak, and Xi Wang, *University of Washington*
- COBRA: Making Transactional Key-Value Stores Verifiably Serializable** 63
Cheng Tan and Changgeng Zhao, *NYU*; Shuai Mu, *Stony Brook University*; Michael Walfish, *NYU*
- Determinizing Crash Behavior with a Verified Snapshot-Consistent Flash Translation Layer** 81
Yun-Sheng Chang, Yao Hsiao, Tzu-Chi Lin, Che-Wei Tsao, Chun-Feng Wu, Yuan-Hao Chang, Hsiang-Shang Ko, and Yu-Fang Chen, *Institute of Information Science, Academia Sinica, Taiwan*
- Storage Systems are Distributed Systems (So Verify Them That Way!)** 99
Travis Hance, *Carnegie Mellon University*; Andrea Lattuada, *ETH Zurich*; Chris Hawblitzel, *Microsoft Research*; Jon Howell and Rob Johnson, *VMware Research*; Bryan Parno, *Carnegie Mellon University*

Storage

- Fast RDMA-based Ordered Key-Value Store using Remote Learned Cache** 117
Xingda Wei, Rong Chen, and Haibo Chen, *Shanghai Jiao Tong University*
- CrossFS: A Cross-layered Direct-Access File System** 137
Yujie Ren, *Rutgers University*; Changwoo Min, *Virginia Tech*; Sudarsun Kannan, *Rutgers University*
- From WiscKey to Bourbon: A Learned Index for Log-Structured Merge Trees** 155
Yifan Dai, Yien Xu, Aishwarya Ganesan, and Ramnathan Alagappan, *University of Wisconsin - Madison*; Brian Kroth, *Microsoft Gray Systems Lab*; Andrea Arpaci-Dusseau and Remzi Arpaci-Dusseau, *University of Wisconsin - Madison*
- LinnOS: Predictability on Unpredictable Flash Storage with a Light Neural Network** 173
Mingzhe Hao, Levent Toksoz, and Nanqinqin Li, *University of Chicago*; Edward Edberg Halim, *Surya University*; Henry Hoffmann and Haryadi S. Gunawi, *University of Chicago*
- A large scale analysis of hundreds of in-memory cache clusters at Twitter** 191
Juncheng Yang, *Carnegie Mellon University*; Yao Yue, *Twitter*; K. V. Rashmi, *Carnegie Mellon University*
- Generalized Sub-Query Fusion for Eliminating Redundant I/O from Big-Data Queries** 209
Partho Sarthi, Kaushik Rajan, and Akash Lal, *Microsoft Research India*; Abhishek Modi, Prakhar Jain, Mo Liu, and Ashit Gosalia, *Microsoft*; Saurabh Kalikar, *Intel*

OS & Networking

- A Simpler and Faster NIC Driver Model for Network Functions** 225
Solal Pirelli and George Candea, *EPFL*
- PANIC: A High-Performance Programmable NIC for Multi-tenant Networks** 243
Jiaxin Lin, *University of Wisconsin - Madison*; Kiran Patel and Brent E. Stephens, *University of Illinois at Chicago*; Anirudh Sivaraman, *New York University (NYU)*; Aditya Akella, *University of Wisconsin - Madison*

Semeru: A Memory-Disaggregated Managed Runtime	261
Chenxi Wang, Haoran Ma, Shi Liu, and Yuanqi Li, <i>UCLA</i> ; Zhenyuan Ruan, <i>MIT</i> ; Khanh Nguyen, <i>Texas A&M University</i> ; Michael D. Bond, <i>Ohio State University</i> ; Ravi Netravali, Miryung Kim, and Guoqing Harry Xu, <i>UCLA</i>	
Caladan: Mitigating Interference at Microsecond Timescales	281
Joshua Fried and Zhenyuan Ruan, <i>MIT CSAIL</i> ; Amy Ousterhout, <i>UC Berkeley</i> ; Adam Belay, <i>MIT CSAIL</i>	
Overload Control for μs-scale RPCs with Breakwater	299
Inho Cho, Ahmed Saeed, Joshua Fried, Seo Jin Park, Mohammad Alizadeh, and Adam Belay, <i>MIT CSAIL</i>	
AIFM: High-Performance, Application-Integrated Far Memory	315
Zhenyuan Ruan, <i>MIT CSAIL</i> ; Malte Schwarzkopf, <i>Brown University</i> ; Marcos K. Aguilera, <i>VMware Research</i> ; Adam Belay, <i>MIT CSAIL</i>	

Consistency

Performance-Optimal Read-Only Transactions	333
Haonan Lu, <i>Princeton University</i> ; Siddhartha Sen, <i>Microsoft Research</i> ; Wyatt Lloyd, <i>Princeton University</i>	
Toward a Generic Fault Tolerance Technique for Partial Network Partitioning	351
Mohammed Alfatafta, Basil Alkhatib, Ahmed Alquraan, and Samer Al-Kiswany, <i>University of Waterloo, Canada</i>	
PACEMAKER: Avoiding HeART attacks in storage clusters with disk-adaptive redundancy	369
Saurabh Kadekodi, Francisco Maturana, Suhas Jayaram Subramanya, Juncheng Yang, K. V. Rashmi, and Gregory R. Ganger, <i>Carnegie Mellon University</i>	
Pegasus: Tolerating Skewed Workloads in Distributed Storage with In-Network Coherence Directories	387
Jialin Li, <i>National University of Singapore</i> ; Jacob Nelson, <i>Microsoft Research</i> ; Ellis Michael, <i>University of Washington</i> ; Xin Jin, <i>Johns Hopkins University</i> ; Dan R. K. Ports, <i>Microsoft Research</i>	
FlightTracker: Consistency across Read-Optimized Online Stores at Facebook	407
Xiao Shi, Scott Pruett, Kevin Doherty, Jinyu Han, Dmitri Petrov, Jim Carrig, John Hugg, and Nathan Bronson, <i>Facebook, Inc.</i>	
KVell+: Snapshot Isolation without Snapshots	425
Baptiste Lepers and Oana Balmau, <i>University of Sydney</i> ; Karan Gupta, <i>Nutanix Inc.</i> ; Willy Zwaenepoel, <i>University of Sydney</i>	

Thursday, November 5

Machine Learning 1

Serving DNNs like Clockwork: Performance Predictability from the Bottom Up	443
Arpan Gujarati, <i>Max Planck Institute for Software Systems</i> ; Reza Karimi, <i>Emory University</i> ; Safya Alzayat, Wei Hao, and Antoine Kaufmann, <i>Max Planck Institute for Software Systems</i> ; Ymir Vigfusson, <i>Emory University</i> ; Jonathan Mace, <i>Max Planck Institute for Software Systems</i>	
A Unified Architecture for Accelerating Distributed DNN Training in Heterogeneous GPU/CPU Clusters	463
Yimin Jiang, <i>Tsinghua University and ByteDance</i> ; Yibo Zhu, <i>ByteDance</i> ; Chang Lan, <i>Google</i> ; Bairen Yi, <i>ByteDance</i> ; Yong Cui, <i>Tsinghua University</i> ; Chuanxiong Guo, <i>ByteDance</i>	
Heterogeneity-Aware Cluster Scheduling Policies for Deep Learning Workloads	481
Deepak Narayanan and Keshav Santhanam, <i>Stanford University and Microsoft Research</i> ; Fiodar Kazhamiaka, <i>Stanford University</i> ; Amar Phanishayee, <i>Microsoft Research</i> ; Matei Zaharia, <i>Stanford University</i>	
PipeSwitch: Fast Pipelined Context Switching for Deep Learning Applications	499
Zhihao Bai and Zhen Zhang, <i>Johns Hopkins University</i> ; Yibo Zhu, <i>ByteDance Inc.</i> ; Xin Jin, <i>Johns Hopkins University</i>	
HiveD: Sharing a GPU Cluster for Deep Learning with Guarantees	515
Hanyu Zhao, <i>Peking University and Microsoft</i> ; Zhenhua Han, <i>The University of Hong Kong and Microsoft</i> ; Zhi Yang, <i>Peking University</i> ; Quanlu Zhang, Fan Yang, Lidong Zhou, and Mao Yang, <i>Microsoft</i> ; Francis C.M. Lau, <i>The University of Hong Kong</i> ; Yuqi Wang, Yifan Xiong, and Bin Wang, <i>Microsoft</i>	
AntMan: Dynamic Scaling on GPU Clusters for Deep Learning	533
Wencong Xiao, Shiru Ren, Yong Li, Yang Zhang, Pengyang Hou, Zhi Li, Yihui Feng, Wei Lin, and Yangqing Jia, <i>Alibaba Group</i>	

Consensus

- Write Dependency Disentanglement with HORAE** 549
Xiaojiang Liao, Youyou Lu, Erci Xu, and Jiwu Shu, *Tsinghua University*
- Blockene: A High-throughput Blockchain Over Mobile Devices** 567
Sambhav Satija and Apurv Mehra, *Microsoft Research India*; Sudheesh Singanamalla, *University of Washington*; Karan Grover, Muthian Sivathanu, Nishanth Chandran, Divya Gupta, and Satya Lokam, *Microsoft Research India*
- Tolerating Slowdowns in Replicated State Machines using Copilots** 583
Khiem Ngo, *Princeton University*; Siddhartha Sen, *Microsoft Research*; Wyatt Lloyd, *Princeton University*
- Microsecond Consensus for Microsecond Applications** 599
Marcos K. Aguilera and Naama Ben-David, *VMware Research*; Rachid Guerraoui, *EPFL*; Virendra J. Marathe, *Oracle Labs*; Athanasios Xygkis and Igor Zlotchi, *EPFL*
- Virtual Consensus in Delos** 617
Mahesh Balakrishnan, Jason Flinn, Chen Shen, Mihir Dharamshi, Ahmed Jafri, Xiao Shi, Santosh Ghosh, Hazem Hassan, Aaryaman Sagar, Rhed Shi, Jingming Liu, Filip Gruszczynski, Xianan Zhang, Huy Hoang, Ahmed Yossef, Francois Richard, and Yee Jiun Song, *Facebook, Inc.*
- Byzantine Ordered Consensus without Byzantine Oligarchy** 633
Yunhao Zhang, *Cornell University*; Srinath Setty, Qi Chen, and Lidong Zhou, *Microsoft Research*; Lorenzo Alvisi, *Cornell University*

Bugs

- From Global to Local Quiescence: Wait-Free Code Patching of Multi-Threaded Processes** 651
Florian Rommel and Christian Dietrich, *Leibniz Universität Hannover*; Daniel Friesel, Marcel Köppen, Christoph Borchert, Michael Müller, and Olaf Spinczyk, *Universität Osnabrück*; Daniel Lohmann, *Leibniz Universität Hannover*
- Testing Database Engines via Pivoted Query Synthesis** 667
Manuel Rigger and Zhendong Su, *ETH Zurich*
- Gauntlet: Finding Bugs in Compilers for Programmable Packet Processing** 683
Fabian Ruffy, Tao Wang, and Anirudh Sivaraman, *New York University*
- Aragog: Scalable Runtime Verification of Shardable Networked Systems** 701
Nofel Yaseen, *University of Pennsylvania*; Behnaz Arzani and Ryan Beckett, *Microsoft Research*; Selim Ciraci, *Microsoft*; Vincent Liu, *University of Pennsylvania*
- Automated Reasoning and Detection of Specious Configuration in Large Systems with Symbolic Execution** 719
Yigong Hu, Gongqi Huang, and Peng Huang, *Johns Hopkins University*
- Testing Configuration Changes in Context to Prevent Production Failures** 735
Xudong Sun, Runxiang Cheng, Jianyan Chen, and Elaine Ang, *University of Illinois at Urbana-Champaign*; Owolabi Legunsen, *Cornell University*; Tianyin Xu, *University of Illinois at Urbana-Champaign*

Scheduling

- Providing SLOs for Resource-Harvesting VMs in Cloud Platforms** 753
Pradeep Ambati, *University of Massachusetts, Amherst*; Íñigo Goiri, Felipe Frujeri, *Microsoft Azure and Microsoft Research*; Alper Gun and Ke Wang, *Google*; Brian Dolan, Brian Corell, Sekhar Pasupuleti, Thomas Moscibroda, Sameh Elnikety, Marcus Fontoura, and Ricardo Bianchini, *Microsoft Azure and Microsoft Research*
- The CacheLib Caching Engine: Design and Experiences at Scale** 769
Benjamin Berg, *Carnegie Mellon University*; Daniel S. Berger, *Carnegie Mellon University and Microsoft Research*; Sara McAllister and Isaac Grosf, *Carnegie Mellon University*; Sathya Gunasekar, Jimmy Lu, Michael Uhlar, and Jim Carrig, *Facebook*; Nathan Beckmann, Mor Harchol-Balter, and Gregory R. Ganger, *Carnegie Mellon University*
- Twine: A Unified Cluster Management System for Shared Infrastructure** 787
Chunqiang Tang, Kenny Yu, Kaushik Veeraraghavan, Jonathan Kaldor, Scott Michelson, Thawan Kooburat, Aravind Anbudurai, Matthew Clark, Kabir Gogia, Long Cheng, Ben Christensen, Alex Gartrell, Maxim Khutorenko, Sachin Kulkarni, Marcin Pawlowski, Tuomas Pelkonen, Andre Rodrigues, Rounak Tibrewal, Vaishnavi Venkatesan, and Peter Zhang, *Facebook Inc.*

FIRM: An Intelligent Fine-grained Resource Management Framework for SLO-Oriented Microservices 805
Haoran Qiu, Subho S. Banerjee, Saurabh Jha, Zbigniew T. Kalbarczyk, and Ravishankar K. Iyer, *University of Illinois at Urbana-Champaign*

Building Scalable and Flexible Cluster Managers Using Declarative Programming 827
Lalith Suresh, *VMware*; João Loff, *IST (ULisboa) / INESC-ID*; Faria Kalim, *UIUC*; Sangeetha Abdu Jyothi, *UC Irvine and VMware*; Nina Narodytska, Leonid Ryzhyk, Sahar Gamage, Brian Oki, Pranshu Jain, and Michael Gasch, *VMware*

Protean: VM Allocation Service at Scale 845
Ori Hadary, Luke Marshall, Ishai Menache, Abhisek Pan, Esaias E Greeff, David Dion, Star Dorminey, Shailesh Joshi, Yang Chen, Mark Russinovich, and Thomas Moscibroda, *Microsoft Azure and Microsoft Research*

Friday, November 6

Machine Learning 2

Ansor: Generating High-Performance Tensor Programs for Deep Learning 863
Lianmin Zheng, *UC Berkeley*; Chengfan Jia, Minmin Sun, and Zhao Wu, *Alibaba Group*; Cody Hao Yu, *Amazon Web Services, Inc*; Ameer Haj-Ali, *UC Berkeley*; Yida Wang, *Amazon Web Services*; Jun Yang, *Alibaba Group*; Danyang Zhuo, *UC Berkeley and Duke University*; Koushik Sen, Joseph E. Gonzalez, and Ion Stoica, *UC Berkeley*

RAMMER: Enabling Holistic Deep Learning Compiler Optimizations with rTasks 881
Lingxiao Ma, *Peking University and Microsoft Research*; Zhiqiang Xie, *ShanghaiTech University and Microsoft Research*; Zhi Yang, *Peking University*; Jilong Xue, Youshan Miao, Wei Cui, Wenxiang Hu, Fan Yang, Lintao Zhang, and Lidong Zhou, *Microsoft Research*

A Tensor Compiler for Unified Machine Learning Prediction Serving 899
Supun Nakandala, *UC San Diego*; Karla Saur, *Microsoft*; Gyeong-In Yu, *Seoul National University*; Konstantinos Karanasos, Carlo Curino, Markus Weimer, and Matteo Interlandi, *Microsoft*

Retiarrii: A Deep Learning Exploratory-Training Framework 919
Quanlu Zhang, Zhenhua Han, Fan Yang, Yuge Zhang, Zhe Liu, Mao Yang, and Lidong Zhou, *Microsoft Research*

KungFu: Making Training in Distributed Machine Learning Adaptive 937
Luo Mai, Guo Li, Marcel Wagenländer, Konstantinos Fertakis, Andrei-Octavian Brabete, and Peter Pietzuch, *Imperial College London*

Hardware

FVM: FPGA-assisted Virtual Device Emulation for Fast, Scalable, and Flexible Storage Virtualization 955
Dongup Kwon, *Department of Electrical and Computer Engineering, Seoul National University / Memory Solutions Lab, Samsung Semiconductor Inc.*; Junehyuk Boo and Dongryeong Kim, *Department of Electrical and Computer Engineering, Seoul National University*; Jangwoo Kim, *Department of Electrical and Computer Engineering, Seoul National University / Memory Solutions Lab, Samsung Semiconductor Inc.*

hXDP: Efficient Software Packet Processing on FPGA NICs 973
Marco Spaziani Brunella and Giacomo Belocchi, *Axbryd/University of Rome Tor Vergata*; Marco Bonola, *Axbryd/CNIT*; Salvatore Pontarelli, *Axbryd*; Giuseppe Siracusano, *NEC Laboratories Europe*; Giuseppe Bianchi, *University of Rome Tor Vergata*; Aniello Cammarano, Alessandro Palumbo, and Luca Petrucci, *CNIT/University of Rome Tor Vergata*; Roberto Bifulco, *NEC Laboratories Europe*

Do OS abstractions make sense on FPGAs? 991
Dario Korolija, Timothy Roscoe, and Gustavo Alonso, *ETH Zurich*

Assise: Performance and Availability via Client-local NVM in a Distributed File System 1011
Thomas E. Anderson, *University of Washington*; Marco Canini, *KAUST*; Jongyul Kim, *KAIST*; Dejan Kostić, *KTH Royal Institute of Technology*; Youngjin Kwon, *KAIST*; Simon Peter, *The University of Texas at Austin*; Waleed Reda, *KTH Royal Institute of Technology and Université catholique de Louvain*; Henry N. Schuh, *University of Washington*; Emmett Witchel, *The University of Texas at Austin*

Persistent State Machines for Recoverable In-memory Storage Systems with NVRam 1029
Wen Zhang, *UC Berkeley*; Scott Shenker, *UC Berkeley/ICSI*; Irene Zhang, *Microsoft Research/University of Washington*

AGAMOTTO: How Persistent is your Persistent Memory Application? 1047
Ian Neal, Ben Reeves, Ben Stoler, and Andi Quinn, *University of Michigan*; Youngjin Kwon, *KAIST*;
Simon Peter, *University of Texas at Austin*; Baris Kasikci, *University of Michigan*

Security

Orchard: Differentially Private Analytics at Scale 1065
Edo Roth, Hengchu Zhang, Andreas Haeberlen, and Benjamin C. Pierce, *University of Pennsylvania*

Achieving 100Gbps Intrusion Prevention on a Single Server 1083
Zhipeng Zhao, Hugo Sadok, Nirav Atre, James C. Hoe, Vyas Sekar, and Justine Sherry, *Carnegie Mellon University*

DORY: An Encrypted Search System with Distributed Trust 1101
Emma Dauterman, Eric Feng, Ellen Luo, Raluca Ada Popa, and Ion Stoica, *University of California, Berkeley*

SafetyPin: Encrypted Backups with Human-Memorable Secrets 1121
Emma Dauterman, *UC Berkeley*; Henry Corrigan-Gibbs, *EPFL and MIT CSAIL*; David Mazières, *Stanford University*

Efficiently Mitigating Transient Execution Attacks using the Unmapped Speculation Contract 1139
Jonathan Behrens, Anton Cao, Cel Skeggs, Adam Belay, M. Frans Kaashoek, and Nickolai Zeldovich, *MIT CSAIL*

Clusters

Predictive and Adaptive Failure Mitigation to Avert Production Cloud VM Interruptions 1155
Sebastien Levy, Randolph Yao, Youjiang Wu, and Yingnong Dang, *Microsoft Azure*; Peng Huang, *Johns Hopkins University*; Zheng Mu, *Microsoft Azure*; Pu Zhao, *Microsoft Research*; Tarun Ramani, Naga Govindaraju, and Xukun Li, *Microsoft Azure*; Qingwei Lin, *Microsoft Research*; Gil Lapid Shafiriri and Murali Chintalapati, *Microsoft Azure*

Sundial: Fault-tolerant Clock Synchronization for Datacenters 1171
Yuliang Li, *Google Inc. and Harvard University*; Gautam Kumar, Hema Hariharan, Hassan Wassel, Peter Hochschild, and Dave Platt, *Google Inc.*; Simon Sabato, *Lilac Cloud*; Minlan Yu, *Harvard University*; Nandita Dukkkipati, Prashant Chandra, and Amin Vahdat, *Google Inc.*

Fault-tolerant and transactional stateful serverless workflows 1187
Haoran Zhang, *University of Pennsylvania*; Adney Cardoza, *Rutgers University–Camden*; Peter Baile Chen, Sebastian Angel, and Vincent Liu, *University of Pennsylvania*

Unearthing inter-job dependencies for better cluster scheduling 1205
Andrew Chung, *Carnegie Mellon University*; Subru Krishnan, Konstantinos Karanasos, and Carlo Curino, *Microsoft*; Gregory R. Ganger, *Carnegie Mellon University*

RackSched: A Microsecond-Scale Scheduler for Rack-Scale Computers 1225
Hang Zhu, *Johns Hopkins University*; Kostis Kaffes, *Stanford University*; Zixu Chen, *Johns Hopkins University*; Zhenming Liu, *College of William and Mary*; Christos Kozyrakis, *Stanford University*; Ion Stoica, *UC Berkeley*; Xin Jin, *Johns Hopkins University*

Thunderbolt: Throughput-Optimized, Quality-of-Service-Aware Power Capping at Scale 1241
Shaohong Li, Xi Wang, Xiao Zhang, Vasileios Kontorinis, Sreekumar Kodakara, David Lo, and Parthasarathy Ranganathan, *Google LLC*