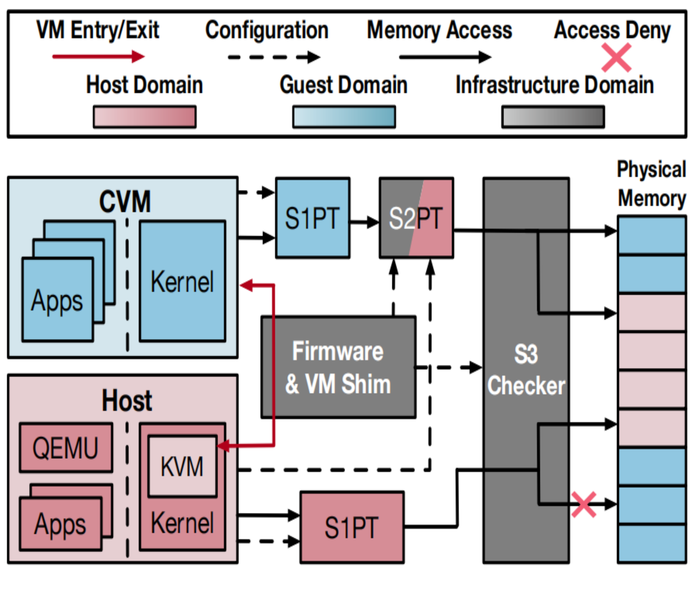
CPC: Flexible, Secure, and Efficient CVM Maintenance with Confidential Procedure Calls

Jiahao Chen, Zeyu Mi, Yubin Xia, Haibing Guan, Haibo Chen (IPADS, INSTITUTE OF PARALLEL AND DISTRIBUTED SYSTEMS, SJTU, China)

Maintenance: Achilles Heel of CVMs



Confidential Virtual Machine (CVM) platforms protect cloud tenants' VM instances in TEEs isolated from the host. (E.g., **AMD SEV**, **Intel TDX**, **ARM CCA,** and **RISC-V CoVE**)

Attestation: Booted images can be verified by the cloud tenant.

State Protection: Guest register states are not directly accessible by the host.

Memory Protection: Stage-3 memory ownership management prevents host access to guest private memory.

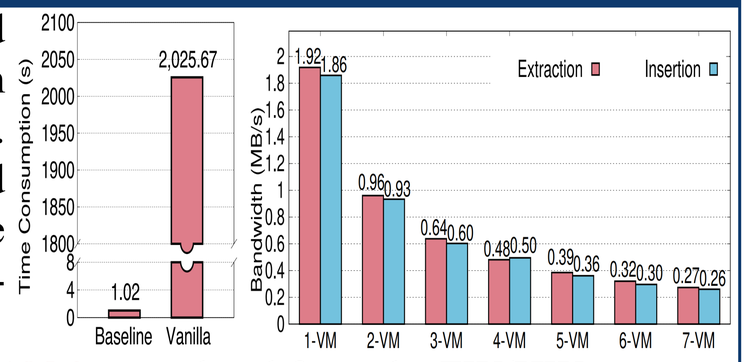
However, CVMs pose significant challenges to VM maintenance…

	Traditional VM	CVM	Use Case
Host-driven	The host directly accesses the internal data of the guests to realize these services.	**Failed!** CVMs block intrusive and direct host accesses to the guest.	(Live) Migration
			Snapshot
			Disaster Recovery
Guest-driven	The tenant installs host's agents in VMs that bridge the host-guest semantic gap and work in conjunction with the host software stack.	**Failed!** Guests deny the untrusted agent software provided by the host.	Logging
			Security Scanning
			Monitoring
			Backup
			Resource Reclamation

Limitations of Current Solutions

Inflexibility: Slow and complicated updates on the hardware/firmware.

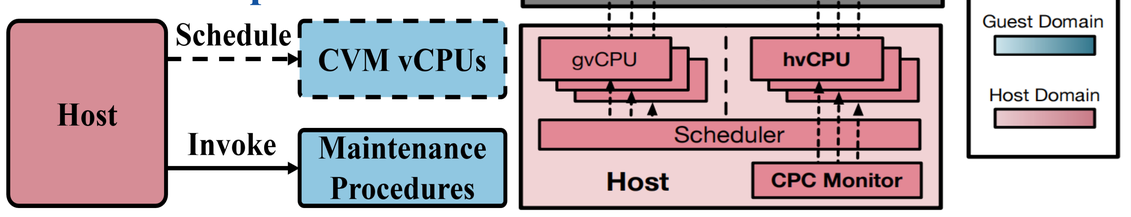
Poor Security: Inflated firmware's code base and complicated host-guest interactions.



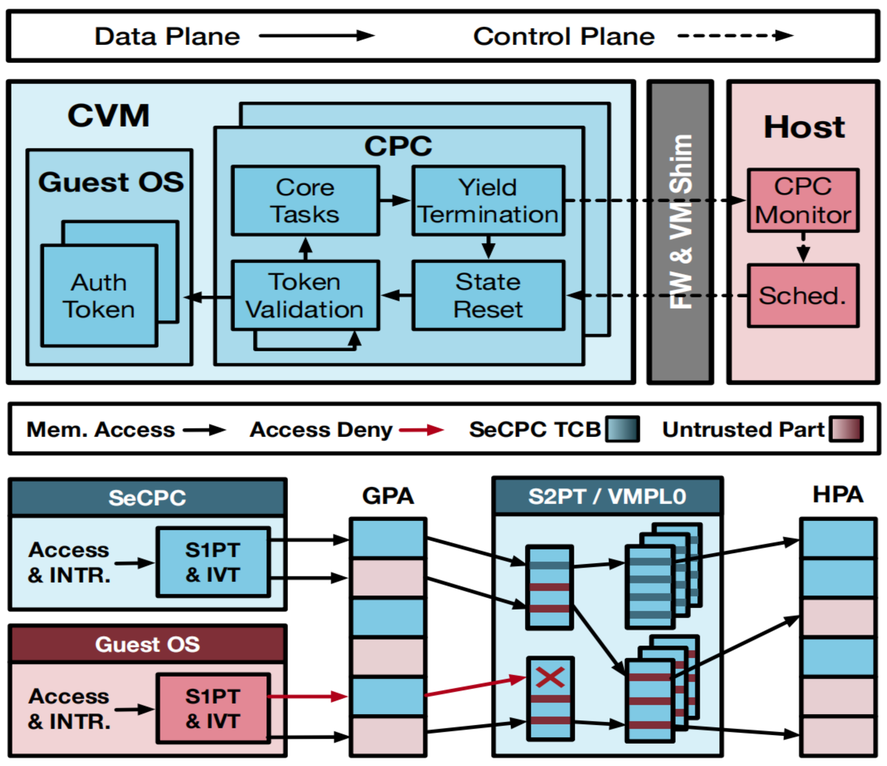
Low Performance: 1) **1986×** overhead from the FW/HW.
2) Interference from guests (**3×** overhead on resource reclamation).

Key Idea & Architecture Overview

Key Idea: Extending the semantic of **host scheduling of vCPUs** to the semantic of **host invocations of CVM maintenance procedures**.



Confidential Procedure Calls (CPCs) with Confidential Page Table Isolation (CPTI)



The core maintenance logic of a CPC is designed as **a state machine** driven by both the **in-host control plane** and the **in-guest data plane**.

- CPCs maintain a clear security boundary between the host and guest, ensuring no security degradation on the CVM.
- The host can precisely invoke the desired maintenance tasks through the semantics bridged by CPCs.

SeCPCs (Secure CPCs) with CPTI can protect security-critical modules even when **the guest OS is compromised**.

The key observation is that: **Although Intel TDX, ARM CCA, and RISC-V CoVE do not have the AMD VMPL feature, they have trusted S2PTs for CVM private memory!**

- **S2PT Isolation: S**eparate stage-2 page tables are established by the trusted firmware for the hvCPUs where SeCPCs reside.
- **S1PT Isolation & IVT Isolation:** Secure stage-1 page tables and interrupt vector tables (IVTs) with secure handlers are established in the isolated memory, which ensures the stable execution of SeCPC's code.

Security Analysis on ARM CCA

Why we like SGX? 🙂 A **c**lear security boundary, only remember "*The outside is bad, the inside is good.*"

Why we want CVM? 😄 SGX has complicated interfaces and host-guest interaction (syscalls) to the host kernel. So we want CVM with simple interfaces and interaction (VM exits).

Why we love CPC? 😍

1) Maintain the clear security boundary

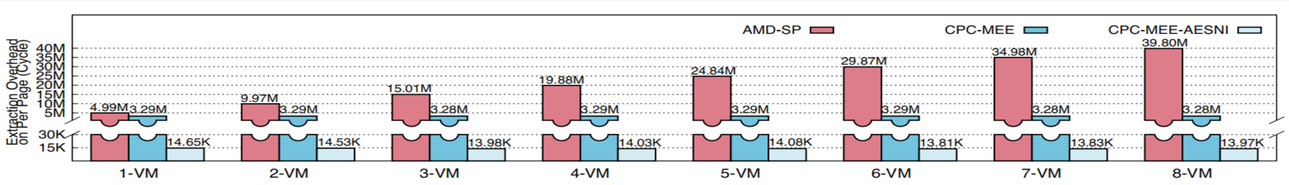
2) Reuse current mature mechanisms and simple interfaces

Compacting Infrastructure Domain: Modifications on the RMM of ARM CCA for CPTI support is **7.23× less** than supporting *CPC-Snapshot* and *CPC-SecureLog* in it.

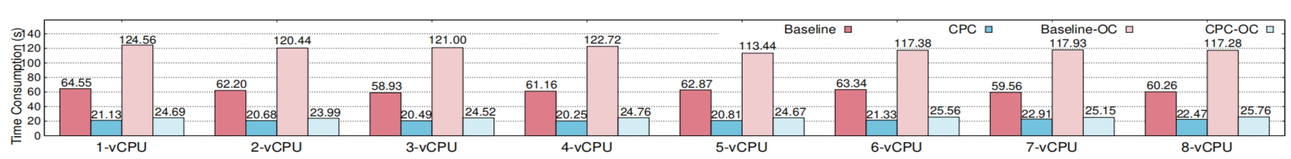
Guest Security: 1) Tiny code base of CPCs (compared with the guest Linux)
2) Timely patches and upgrades
3) Defending compromised CPCs by CPTI
4) Equipping only the needed CPCs

Host Security: Few modifications on the host Linux/KVM, only forwarding the CPC-related hypercalls to the user-level VMM and providing additional memory to the RMM for CPTI. Most modifications are in QEMU/KVMTOOL.

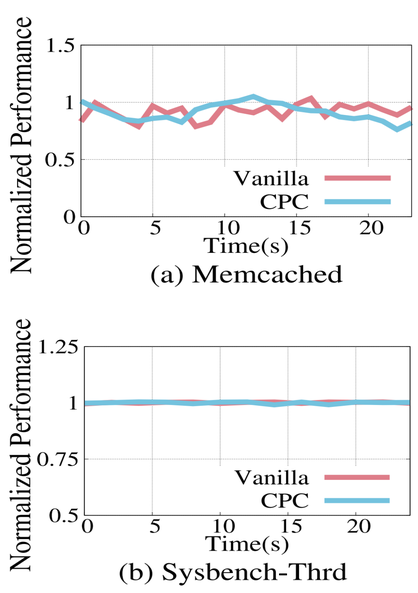
Performance Evaluation on AMD SEV

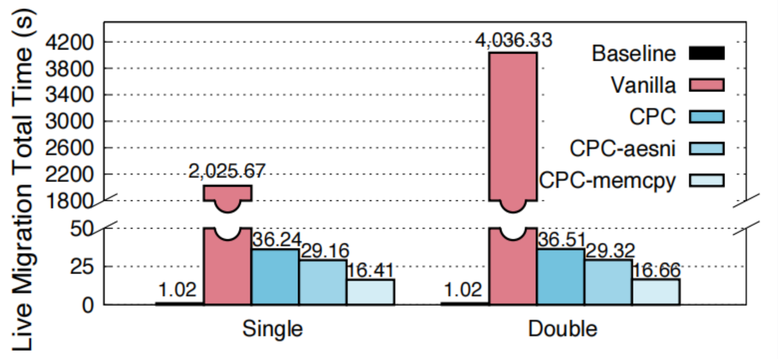


CPC-Snapshot: **340.61× faster** than current SEV CVMs, and can be **2800× or more** when CVM instances increase.



CPC-Ballooning: **3.05× faster** than the baseline, and can be **5× or more** when cloud vendors overcommit vCPUs.





CPC-LiveMigration: **69.47× faster** than the AMD-SP solution, and can be **137.66× or more** when CVM instances increase.

No significant performance degradation on the guest workloads was observed (**<9%**).

Contact Us

Email: chenjiahaosys@gmail.com

WeChat

My CPC talk will be on July 12 around 11:30 on ATC24 track2. I will also present another paper called Jiagu on July 10 around 10:30 p.m. at ATC24 track1 for my lab mates. Looking forward to discuss with you then!

You can also contact this Email address to ask more questions about Jiagu: liu_qy@sjtu.edu.cn, and the poster is below.

Paper Timeline of OSDI/ATC'24

I also made a timeline below for the three tracks of OSDI24 and ATC24. Maybe that would help you.

PS: The paper titles maybe changed in their final versions.

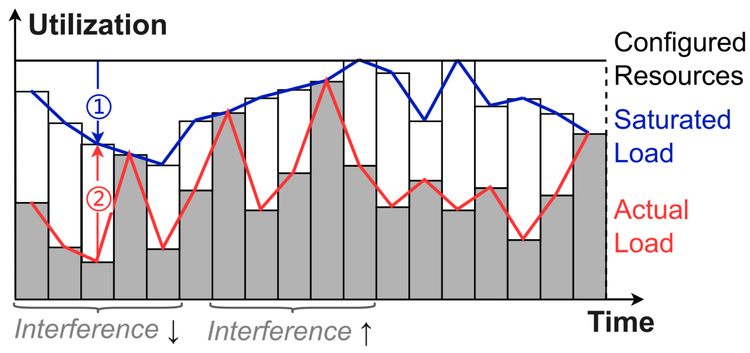
You can find more goodies below~

IPADS, SJTU, China http://ipads.se.sjtu.edu.cn

Harmonizing Efficiency and Practicability: Optimizing Resource Utilization in Serverless Computing with Jiagu

Qingyuan Liu, Yanning Yang, Dong Du, Yubin Xia, Ping Zhang, Jia Feng,
James R. Larus, Haibo Chen

Motivation



Two causes of resource wastage:

- Part I: resource overprovisioning
- Part II: unpredictable load fluctuation

Reduce wastage: current methods and challenges (tradeoffs)

Part I: Overcommitment

- Predict QoS violation when scheduling
- Deploy more instances while ensuring QoS
- **Can't achieve both accuracy & low cost**

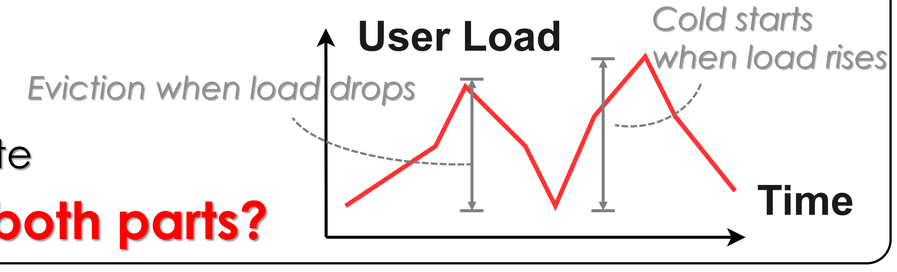
Complex model: accurate but slow

Heuristic model: Fast but inaccurate

How to break the tradeoffs of both parts?

Part II: Autoscaling

- Dynamically create/evict instances upon load fluctuations
- Scaling with higher sensitivity: **higher utilization** but **more cold starts**



Design I: Pre-decision Scheduling

QoS prediction:

$$P_{A\cup\{B,C,...\}} = RFR\{P_A, R_A, C_A, R_B, C_B, R_C, C_C, ...\}$$

Features of co-located functions

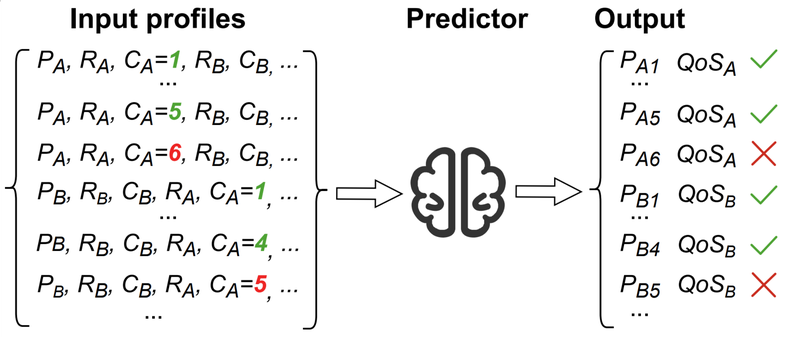
Solo-run performance *Profiles* *Concurrency*

Insight: decouple prediction and decision making

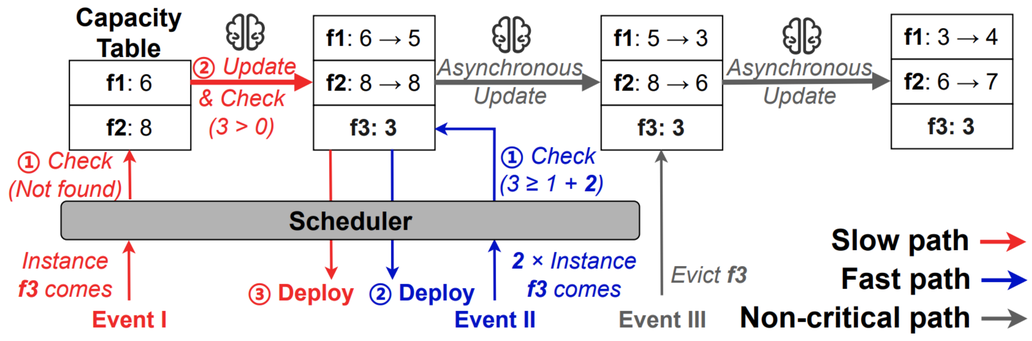
Predict the capacity for each function on a server

- Capacity: the maximum QoS-guaranteed concurrency
- Inspired by Serverless unique feature: more replicated instances of deployed instances are likely to arrive later

Calculate the capacity:



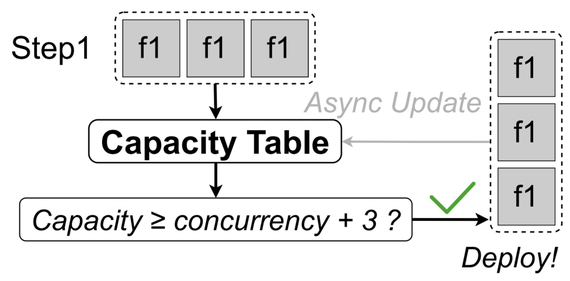
Example:



Goal: accurate prediction with practical cost

Scheduling Fast path:

- Condition: if the arriving instance matches a capacity table entry
- Scheduling by checking the capacity (fast)
- Concurrent schedulings can be batched



Scheduling Slow path:

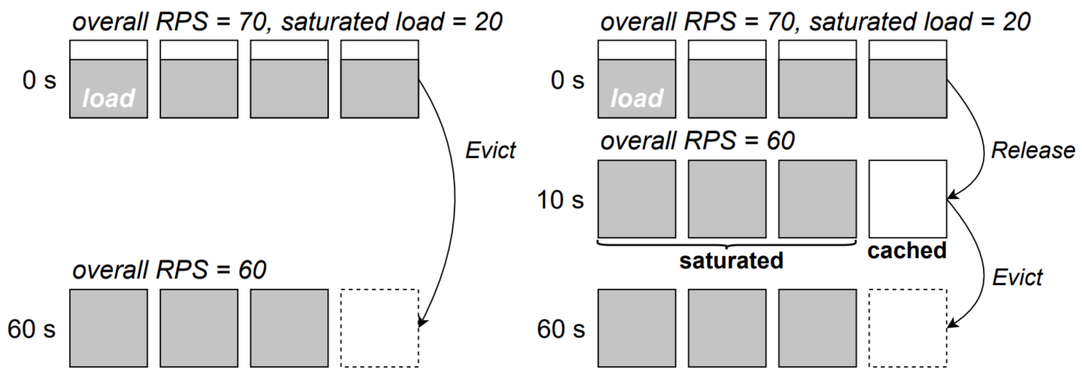
- No matched capacity table entry
- Need to calculate the capacity before deployment

Asynchronous update:

- Update the capacity table after deployment
- Remove costly schedulings from the critical path

Design II: Dual-staged Scaling

Insight: decouple resource allocation/release and instance creation/eviction



(a) Existing eviction. (b) Dual-staged eviction.

Goal: sensitive autoscaling without additional cold starts

Add a release stage between eviction

- Cached instances do not serve requests: release resources by re-routing **(<1ms)**
- Higher sensitivity than eviction ⇒ higher resource utilization

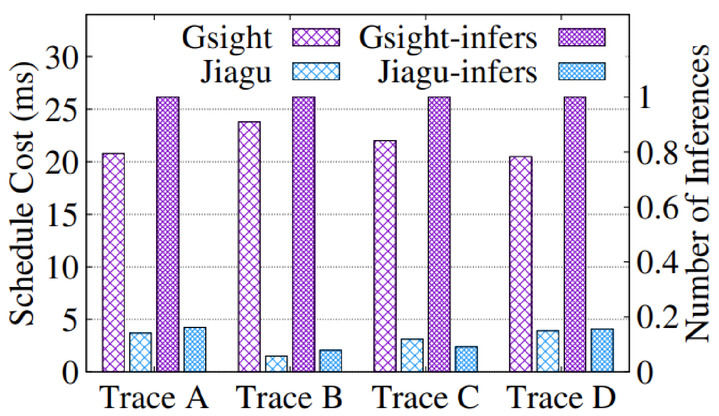
Logical cold start: if the load re-rises before eviction

- Re-route: convert cached instances to saturated instances
- Negligible cost **(<1ms)**

On-demand migration

- Migrate cached instances in advance if the server is full

Evaluation

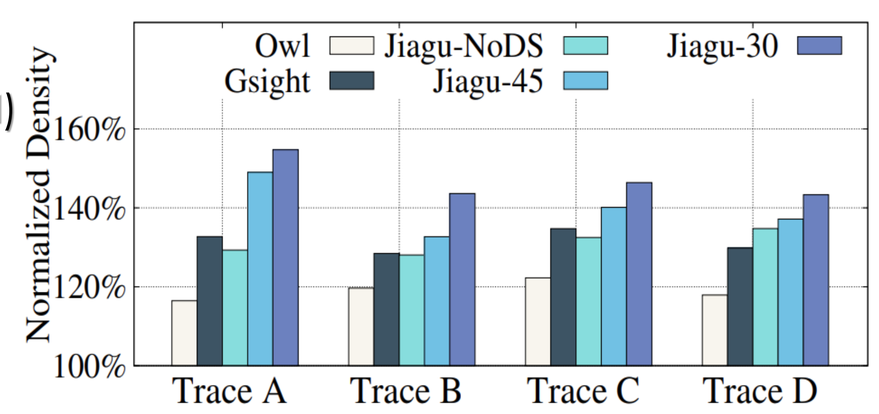


Reduce Cost

- **81.0%–93.7%** lower scheduling costs
- Less number of cold starts (by Design II) & less scheduling costs (by Design I)

Effective Scheduling

- **22%** higher deployment density than Gsight
- QoS meets the goal (<10% violation)



Wed, July 10	OSDI '24	USENIX ATC '24	USENIX ATC '24
8:00 am–9:00 am	Continental Breakfast		
9:00 am–10:00 am	OSDI '24 and USENIX ATC '24 Joint Keynote Address Scaling AI Sustainably: An Uncharted Territory Carole-Jean Wu, Meta		
10:00 am–10:30 am	Break with Refreshments		
10:30 am–10:45 am	Opening Remarks and Awards (Program Co-Chairs: Ada Gavrilovska, Georgia Institute of Technology; Douglas B. Terry, Amazon Web Services)	Opening Remarks Saurabh Bagchi, Purdue University; Yiying Zhang, University of California, San Diego	
10:45 am–12:45 pm	Memory Management **Sabre: Improving Memory Prefetching in Serverless MicroVMs with Near-Memory Hardware-Accelerated Compression** Nikita Lazarev and Varun Gohil, MIT; James Tsai, Intel Labs; Andy Anderson and Bhushan Chitlur, Intel; Zhiru Zhang, Cornell University; Christina Delimitrou, MIT **Nomad: Non-Exclusive Memory Tiering via Transactional Page Migration** 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** Yuhong Zhong, Columbia University; Daniel S. Berger, Microsoft Azure, University of Washington, and CMU; Carl Waldspurger, Carl Waldspurger Consulting; Ishwar Agarwal, Rajat Agarwal, Frank Hady, and Karthik Kumar, Intel; Mark D. Hill, Microsoft Azure; Mosharaf Chowdhury, University of Michigan; Asaf Cidon, Columbia University **Harvesting Memory-bound CPU Stall Cycles in Software with MSH** Zhihong Luo, Sam Son, and Sylvia Ratnasamy, UC Berkeley; Scott Shenker, ICSI and UC Berkeley **A Tale of Two Paths: Toward a Hybrid Data Plane for Efficient Far-Memory Applications** Lei Chen, Chinese Academy of Sciences; Shi Liu, UCLA; Chenxi Wang, Chinese Academy of Sciences; Haoran Ma and Yifan Qiao, UCLA; Zhe Wang and Chenggang Wu, Chinese Academy of Sciences; Youyou Lu, Tsinghua University; Xiaobing Feng and Huimin Cui, 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** 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	Cloud Computing (I will replace Qingyuan Liu for this talk. My own talk is on July 12.) **Harmonizing Efficiency and Practicability: Optimizing Resource Utilization in Serverless Computing with Jiagu** Qingyuan Liu, Yanning Yang, Dong Du, and Yubin Xia, Shanghai Jiao Tong University; Ping Zhang and Jia Feng, Huawei Cloud; James Larus, EPFL; Haibo Chen, Shanghai Jiao Tong University **SEALS: A Self-Adaptive, Learned Scheduler for Serverless Functions** Yuqi Fu, University of Virginia; Ruizhe Shi, George Mason University; Haoliang Wang, Adobe Research; Songqing Chen, George Mason University; Yue Cheng, University of Virginia **Starburst: A Cost-aware Scheduler for Cloud Bursting** Michael Luo, Suryaprakash Vengadesan, Siyuan Zhuang, and Romil Bhardwaj, UC Berkeley; Justin Chang, UCSB; Eric Friedman, ICSI and UC Berkeley; Scott Shenker, ICSI and UC Berkeley; Ion Stoica, UC Berkeley **StreamBox: A Lightweight GPU SandBox for Serverless Inference Workflow** Hao Wu, Yue Yu, and Junxiao Deng, Huazhong University of Science and Technology; Shadi Ibrahim, Inria; Ziyue Cheng, Hao Fan, Song Wu, and Hai Jin, Huazhong University of Science and Technology 12:25 end	ML Inference **Power-aware Deep Learning Model Serving with μ-Serve** Haoran Qiu, Weichao Mao, Archit Patke, and Shengkun Cui, University of Illinois Urbana-Champaign; Saurabh Jha, Chen Wang, and Hubertus Franke, IBM Research; Zbigniew Kalbarczyk, Tamer Başar, and Ravishankar K. Iyer, University of Illinois Urbana-Champaign **Fast Inference for Probabilistic Graphical Models** Jiantong Jiang, The University of Western Australia; Zeyi Wen, The Hong Kong University of Science and Technology (Guangzhou); Atif Mansoor and Ajmal Mian, The University of Western Australia **Cost-Efficient Large Language Model Serving for Multi-turn Conversations with CachedAttention** Bin Gao, National University of Singapore; Zhuomin He, Shanghai Jiaotong University; Puru Sharma, Qingxuan Kang, and Djordje Jevdjic, National University of Singapore; Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo, Huawei Cloud **PUZZLE: Efficiently Aligning Large Language Models through Light-Weight Context Switch** Kinman Lei, Mingshu Zhai, Yuyang Jin, Kezhao Huang, Haoxing Ye, and Jidong Zhai, Tsinghua University
12:45 pm–2:00 pm	Symposium Luncheon Sponsored by Roblox		
2:00 pm–3:40 pm	Low-Latency LLM Serving **StableGen: Efficient LLM Inference with Low Tail Latency** Amey Agrawal, Georgia Institute of Technology; Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, and Bhargav Gulavani, Microsoft Research; Alexey Tumanov, Georgia Institute of Technology; Ramachandran Ramjee, Microsoft Research **Phantom: Low-Latency Serverless Inference for Large Language Models** 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** Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim, Seoul National University **Llumnix: Dynamic Scheduling for Large Language Model Serving** Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin, Alibaba Group **DistLLM: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving** Yinmin Zhong and Shengyu Liu, Peking University; Junda Chen, UCSD; Jianbo Hu, Peking University; Yibo Zhu, unaffiliated; Xuanzhe Liu and Xin Jin, Peking University; Hao Zhang, UCSD	Storage 1 **ScalaAFA: Constructing User-Space All-Flash Array Engine with Holistic Designs** Shushu Yi and Xiurui Pan, Peking University; Qiao Li, Xiamen University; Qiang Li, Alibaba; Chenxi Wang, Chinese Academy of Sciences; Bo Mao, Xiamen University; Myoungsoo Jung, KAIST and Panmnesia; Jie Zhang, Peking University **XCommit: resource-efficient, performant and cost-effective file system journaling** Harshad Shirwadkar, Saurabh Kadekodi, and Theodore Ts'o, Google **ZMS: Zone Abstracton for Mobile Flash Storage** Joo-Young Hwang, Seokhwan Kim, Daejun Park, Yong-Gil Song, Junyoung Han, Seunghyun Choi, and Sangyeun Cho, Samsung Electronics; Youjip Won, Korea Advanced Institute of Science and Technology (KAIST) **Ethane: An Asymmetric File System for Disaggregated Persistent Memory** Miao Cai and Junru Shen, College of Computer Science and Software Engineering, Hohai University; Baoliu Ye, State Key Laboratory for Novel Software Technology, Nanjing University, and College of Computer Science and Software Engineering, Hohai University	Networks 1 **PeRF: Preemption-enabled RDMA Framework** Sugi Lee and Mingyu Choi, Acryl Inc.; Ikjun Yeom and Younghoon Kim, Sungkyunkwan University **CyberStar: Simple, Elastic and Cost-Effective Network Functions Management in Cloud Network at Scale** Tingting Xu, unaffiliated; Shunmin Zhu , Song Yang, Xiaomin Wu, Zhigang Zong, Xiaoxin Peng, Bengbeng Xue, Botao Yan, Yilong Lv, Alibaba Group; Camtu Nguyen Xiaoliang Wang, Nanjing University **OSMOSIS: Enabling Multi-Tenancy in Datacenter SmartNICs** Mikhail Khalilov, Marcin Andrzej Chrapek, Siyuan Shen, Thomas Benz, Alessandro Vezzu, Salvatore Di Girolamo, and Timo Schneider, ETH Zurich Daniele De Sensi, Sapienza University of Rome; Luca Benini and Torsten Hoefler, ETH Zurich **ETC: An Elastic Transmission Control Using End-to-End Available Bandwidth Perception** Feixue Han, Tsinghua Shenzhen International Graduate School; Qing Li, Peng Cheng Laboratory; Peng Zhang, Tencent; Gareth Tyson, Hong Kong University of Science and Technology; Yong Jiang, Tsinghua Shenzhen International Graduate School; Mingwei Xu, Tsinghua University; Yulong Lan and ZhiCheng Li, Tencent
3:40 pm–4:10 pm	Break with Refreshments		
4:10 pm–5:40 pm	Distributed Systems **Anon : an FPGA-based Collective Engine for Distributed Applications** Zhenhao He, Dario Korolija, Yu Zhu, and Benjamin Ramhorst, ETH Zurich; Tristan Laan, University of Amsterdam; Lucian Petrica and Michaela Blott, Research Labs AMD Xilinx; Gustavo Alonso, ETH Zurich **Beaver: Practical Partial Snapshots for Distributed Cloud Services** Liangcheng Yu, University of Pennsylvania; Xiao Zhang, Shanghai Jiao Tong University; Haoran Zhang, University of Pennsylvania; John Sonchack, Princeton University; Dan Ports, Microsoft and University of Washington; Vincent Liu, University of Pennsylvania **Fast and Scalable In-network Lock Management Using Lock Fission** Hanze Zhang, Ke Cheng, Rong Chen, and Haibo Chen, Shanghai Jiao Tong University **Chop Chop: Byzantine Atomic Broadcast to the Network Limit** Martina Camaioni, Rachid Guerraoui, Matteo Monti, Pierre-Louis Roman, Manuel Vidigueira, and Gauthier Voron, EPFL 5:30 end	Edge Computing **More is Different: Prototyping and Analyzing a New Form of Edge Server with Massive Mobile SoCs** Li Zhang, Beijing University of Posts and Telecommunications; Zhe Fu, Tsinghua University; Boqing Shi and Xiang Li, Beijing University of Posts and Telecommunications; Rujin Lai and Chenyang Yang, vclusters; Ao Zhou, Xiao Ma, Shangguang Wang, and Mengwei Xu, Beijing University of Posts and Telecommunications **HiP4-UPF: Towards High-Performance Comprehensive 5G User Plane Function on P4 Progammable Switches** Zhixin Wen and Guanhua Yan, Binghamton University, State University of New York **KEPC-Push: A Knowledge-Enhanced Proactive Content Push Strategy for Edge-Assisted Video Feed Streaming** Ziwen Ye, Tsinghua University; Qing Li, Peng Cheng Laboratory; Chunyu Qiao, ByteDance; Xiaoteng Ma, Tsinghua University; Yong Jiang, Tsinghua Shenzhen International Graduate School; Qian Ma and Shengbin Meng, ByteDance; Zhenhui Yuan, University of Warwick; Zili Meng, HKUST **High-density Mobile Cloud Gaming on Edge SoC Farms** Li Zhang, Shangguang Wang, and Mengwei Xu, Beijing University of Posts and Telecommunications	Operating Systems 1 **Opportunities and Limitations of Modern Hardware Isolation Mechanisms** Xiangdong Chen and Zhaofeng Li, University of Utah; Tirth Jain, Birla Institute of Technology and Science, Pilani; Vikram Narayanan and Anton Burtsev, University of Utah **FetchBPF: Customizable Prefetching Policies in Linux with eBPF** Xuechun Cao, Shaurya Patel, and Soo Yee Lim, University of British Columbia; Xueyuan Han, Wake Forest University; Thomas Pasquier, University of British Columbia **Fast (Trapless) Kernel Probes Everywhere** Jinghao Jia, University of Illinois Urbana-Champaign; Michael Le, IBM Research; Salman Ahmed, IBM Research, Yorktown Heights; Dan Williams, Virginia Tech; Hani Jamjoom, IBM; Tianyin Xu, University of Illinois at Urbana-Champaign **HydraRPC: RPC in the CXL Era** Teng Ma, Alibaba Group; Zheng Liu, Zhejiang University and Alibaba Group; Chengkun Wei, Zhejiang University; Jialiang Huang, Tsinghua University; Youwei Zhuo, Alibaba Group; Haoyu Li, Zhejiang University; Ning Zhang, Yijin Guan, and Dimin Niu, Alibaba Group; Mingxing Zhang, Tsinghua University; Tao Ma, Alibaba Group **Enabling Application-Aware Memory Page Placement Policies and Mechanisms With ExtMem** Sepehr Jalalian, Shaurya Patel, Milad Rezaei Hajidehi, and Margo Seltzer, University of British Columbia; Alexandra (Sasha) Fedorova, University of British Columbia and MongoDB
6:00 pm–7:30 pm	OSDI '24 Poster Session and Reception Sponsored by Amazon		

Thur, July 11	OSDI '24	USENIX ATC '24	USENIX ATC '24
8:00 am–9:00 am	Continental Breakfast		
9:00 am–10:40 am	Deep Learning **Enabling Tensor Language Model to Assist in Generating High-Performance Tensor Programs for Deep Learning** 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 **Bitter: Enabling Efficient Low-Precision Deep Learning Computing through Hardware-aware Tensor Transformation** Lei Wang, 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** Qizheng Zhang, Stanford University; Ali Imran, Purdue University; Enkeleda Bardhi, Sapienza University of Rome; Tushar Swamy, unaffiliated; Muhammad Shahbaz, Purdue University; Kunle Olukotun, Stanford University **Cuber: Constraint-Guided Parallelization Plan Generation for Deep Learning Training** Zhiqi Lin, USTC; Youshan Miao, Quanlu Zhang, Fan Yang, and Yi Zhu, Microsoft Research; Cheng Li, USTC; Saeed Maleki, Xu Cao, Ning Shang, Yilei Yang, Weijiang Xu, and Mao Yang, Microsoft Research; Lintao Zhang, BaseBit Technologies; Lidong Zhou, Microsoft Research **Automatic and Efficient Customization of Neural Networks for ML Applications** Yuhan Liu, University of Chicago; Chengcheng Wan, The University of Chicago; Kuntai Du, Henry Hoffmann, and Junchen Jiang, University of Chicago; Shan Lu, Microsoft Research / University of Chicago; Michael Maire, The University of Chicago	Operating Systems 2 **TeleScale: Telemetry for Gargantuan Memory Footprint Applications** Alan Nair, The University of Edinburgh; Sandeep Kumar and Aravinda Prasad, Intel Labs; Ying Huang, Intel; Andy Rudoff, Intel Corporation; Sreenivas Subramoney, Intel **An Empirical Study of Rust-for-Linux: The Success, Dissatisfaction, and Compromise** Hongyu Li, Beijing University of Posts and Telecommunications; Liwei Guo, University of Electronic Science and Technology of China; Yexuan Yang, Shangguang Wang, and Mengwei Xu, Beijing University of Posts and Telecommunications **Scalable and Effective Page-table and TLB management on NUMA Systems** Bin Gao, Qingxuan Kang, and Hao-Wei Tee, National University of Singapore; Kyle Timothy Ng Chu, Horizon Quantum Computing; Alireza Sanaee, Queen Mary University of London; Djordje Jevdjic, National University of Singapore **UniMem: Redesigning Disaggregated Memory within A Unified Local-Remote Memory Hierarchy** Yijie Zhong, Minqiang Zhou, and Zhirong Shen, Xiamen University	Correctness **WingFuzz: Implementing Continuous Fuzzing for DBMSs** Jie Liang, Zhiyong Wu, and Jingzhou Fu, Tsinghua University; Yiyuan Bai and Qiang Zhang, Shuimu Yulin Technology Co., Ltd.; Yu Jiang, Tsinghua University **Balancing Analysis Time and Bug Detection: Daily Development-friendly Bug Detection in Linux** Keita Suzuki, Keio University; Kenta Ishiguro, Hosei University; Kenji Kono, Keio University **Koncord: Verifying Cluster Management Systems** Bingzhe Liu and Gangmuk Lim, UIUC; Ryan Beckett, Microsoft Research; Brighten Godfrey, UIUC and VMware **Monarch: A Fuzzing Framework for Distributed File Systems** Tao Lyu, EPFL; Liyi Zhang, University of Waterloo; Zhiyao Feng, Yueyang Pan, and Yujie Ren, EPFL; Meng Xu, University of Waterloo; Mathias Payer and Sanidhya Kashyap, EPFL
10:40 am–11:10 am	Break with Refreshments		
11:10 am–12:50 pm	Operating Systems **SquirrelFS: using the Rust compiler to check file-system crash consistency** Hayley LeBlanc, Nathan Taylor, James Bornholt, and Vijay Chidambaram, University of Texas at Austin **High-throughput and Flexible Host Networking via Control and Data Path Physical Separation** 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 **PEOS: Persistent Embedded Operating System and Language Support for Multi-threaded Intermittent Computing** 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** Ao Li and Ning Zhang, Washington University in St. Louis **Microkernel Goes General: Performance and Compatibility in the GMK Production Microkernel** Haibo Chen, Huawei Technologies Co., Ltd 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 Technologies Co., Ltd	ML Training **Accelerating the Training of Large Language Models using Efficient Activation Rematerialization and Optimal Hybrid Parallelism** Tailing Yuan, Yuliang Liu, Xucheng Ye, Shenglong Zhang, Jianchao Tan, Bin Chen, Chengru Song, and Di Zhang, Kuaishou Technology **Metis: Fast Automatic Distributed Training on Heterogeneous GPUs** Taegeon Um, Byungsoo Oh, Minyoung Kang, Woo-Yeon Lee, Goeun Kim, Dongseob Kim, Youngtaek Kim, and Mohd Muzzammil, Samsung Research; Myeongjae Jeon, UNIST **FwdFL: Efficient Federated Finetuning of Language Models** Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang, Beijing University of Posts and Telecommunications 12:25 end	Security 1 **A Secure, Fast, and Resource-Efficient Serverless Platform with Function REWIND** Jaehyun Song, Sungkyunkwan University; Bumsuk Kim, Samsung Electronics; Minwoo Kwak, Yonsei University; Byoungyoung Lee, Seoul National University; Euiseong Seo, Sungkyunkwan University; Jinkyu Jeong, Yonsei University **SimEnc: A High-Performance Similarity-Preserving Encryption Approach for Deduplication of Encrypted Docker Images** Tong Sun and Bowen Jiang, Zhejiang University; Borui Li, Southeast University; Jiamei Lv, Yi Gao, and Wei Dong, Zhejiang University **mmTLS: Scaling the Performance of Encrypted Network Traffic Inspection** Junghan Yoon, Seunghyun Do, and Duckwoo Kim, KAIST; Taejoong Chung, Virginia Tech; KyoungSoo Park, KAIST
12:50 pm–2:00 pm	Symposium Luncheon		
2:00 pm–3:40 pm	Cloud Computing **What will it take for Johnny to know when his Cloud job will finish? Towards providing reliable job completion time predictions using PCS** Abdullah Bin Faisal, Noah Martin, Hafiz Mohsin Bashir, Swaminathan Lamelas, and Fahad Dogar, Tufts University **Optimizing Resource Allocation in Hyperscale Datacenters: Scalability, Usability, and Experiences** Neeraj Kumar, Pol Mauri Ruiz, Vijay Menon, Igor Kabiljo, Mayank Pundir, Andrew Newell, Daniel Lee, Liyuan Wang, and Chunqiang Tang, Meta **μSlope: High Compression and Fast Search on Semi-Structured Logs** Rui Wang, YScope; Devin Gibson, University of Toronto and YScope; Kirk Rodrigues, YScope; Yu Luo, YScope, Uber, and University of Toronto; Yun Zhang, Kaibo Wang, Yupeng Fu, and Ting Chen, Uber; Ding Yuan, University of Toronto and YScope **ServiceLab: Detecting Tiny Performance Regressions at Hyperscale** Mike Chow, Meta; Yang Wang, The Ohio State University and Meta; 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, Elvis Huang, and Chunqiang Tang, Meta **Dynamic Scheduling of ML Training across Geo-Distributed Datacenters: Principles and Experiences** Arnab Choudhury, Meta; Yang Wang, The Ohio State University and Meta; Tuomas Pelkonen, Rentle; Kutta Srinivasan, LinkedIn; Abha Jain, Shenghao Lin, Delia David, Siavash Soleimanifard, Michael Chen, Abhishek Yadav, Ritesh Tijoriwala, and Chunqiang Tang, Meta	ML-System Co-Design **Cost-Efficient Machine Learning Input Data Preprocessing** Oto Mraz, Dan-Ovidiu Graur, Muyu Li, and Sepehr Pourghannad, ETH Zurich; Chandramohan A. Thekkath, Google; Ana Klimovic, ETH Zurich **OPER: Optimality-Guided Embedding Table Parallelization for Large-scale Recommendation Model** Zheng Wang, University of California, San Diego; Yuke Wang, Boyuan Feng, and Guyue Huang, University of California, Santa Barbara; Dheevatsa Mudigere and Bharath Muthiah, Meta; Ang Li, Pacific Northwest National Laboratory; Yufei Ding, University of California, San Diego **DeepVisor: Effective Operator Graph Instantiation for Deep Learning by Execution State Monitoring** Chen Zhang, Rongchao Dong, Haojie Wang, Runxin Zhong, Jike Chen, and Jidong Zhai, Tsinghua University **Quant-LLM: Accelerating the Serving of Large Language Models via FP6-Centric Algorithm-System Co-Design on Modern GPUs** Haojun Xia, University of Sydney; Zhen Zheng and Xiaoxia Wu, Microsoft; Shiyang Chen, Rutgers University; Zhewei Yao, Stephen Youn, Arash Bakhtiari, and Michael Wyatt, Microsoft; Donglin Zhuang and Zhongzhu Zhou, University of Sydney; Olatunji Ruwase, Yuxiong He, and Shuaiwen Leon Song, Microsoft	Networks 2 **QDSR: Accelerating Layer-7 Load Balancing by Direct Server Return with QUIC** Ziqi Wei, Tsinghua University; Zhiqiang Wang, Tencent; Qing Li, Peng Cheng Laboratory; Yuan Yang, Tsinghua University; Cheng Luo and Fuyu Wang, Tencent; Yong Jiang, Tsinghua Shenzhen International Graduate School; Sijie Yang, Tencent; Zhenhui Yuan, Northumbria University **Evaluating Chiplet-based Large-Scale Interconnection Networks via Cycle-Accurate Packet-Parallel Simulation** Yinxiao Feng and Yuchen Wei, Institute for Interdisciplinary Information Sciences, Tsinghua University; Dong Xiang, School of Software, Tsinghua University; Kaisheng Ma, Institute for Interdisciplinary Information Sciences, Tsinghua University **Config-Snob: Tuning for the Best Configurations of Networking Protocol Stack** Manaf Bin-Yahya, Yifei Zhao, Hossein Shafieirad, and Anthony Ho, Huawei Technologies Canada; Shijun Yin, Fanzhao Wang, and Geng Li, Huawei Technologies China **Conspirator: SmartNIC-Aided Control Plane for Distributed ML Workloads** Yunming Xiao, Northwestern University; Diman Zad Tootaghat, Aditya Dhakal, Lianjie Cao, and Puneet Sharma, Hewlett Packard Labs; Aleksandar Kuzmanovic, Northwestern University
3:40 pm–4:10 pm	Break with Refreshments		
4:10 pm–5:50 pm	Formal Verification **Automatically Reasoning About How Systems Code Uses the CPU Cache** Rishabh Iyer, Katerina Argyraki, and George Candea, EPFL **VeriSMo: A Verified Security Module for Confidential VMs** Ziqiao Zhou, Microsoft Research; Anjali, University of Wisconsin-Madison; Weiteng Chen, Microsoft Research; Sishuai Gong, Purdue University; Chris Hawblitzel, Microsoft; Weidong Cui, Microsoft Research **Validating the eBPF Verifier via State Embedding** Hao Sun and Zhendong Su, ETH Zurich **Using Dynamically Layered Definite Releases for Verifying the RefFS File System** Mo Zou and Dong Du, Shanghai Jiao Tong University; Mingkai Dong, Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University; Haibo Chen, Shanghai Jiao Tong University **Anvil: Verifying Liveness of Cluster Management Controllers** 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	Memory **Making Memory Management Extensible With Filesystems** Bijan Tabatabai, University of Wisconsin—Madison; James Sorenson and Michael Swift, University of Wisconsin—Madison **Mangosteen: Fast Transparent Durability for Linearizable Applications using NVM** Sergey Egorov, Gregory Chockler, and Brijesh Dongol, University of Surrey; Dan O'Keeffe, Royal Holloway, University of London; Sadegh Keshavarzi, University of Surrey **FlexMem: Adaptive Page Profiling and Migration for Tiered Memory** Dong Xu, University of California, Merced; Junhee Ryu, Jinho Baek, and Kwangsik Shin, SK Hynix; Pengfei Su and Dong Li, University of California, Merced 5:25 end	Reliability **Ammit: Improving Cloud AI Infrastructure Reliability with Proactive Validation** Yifan Xiong, Yuting Jiang, Ziyue Yang, and Lei Qu, Microsoft Research; Guoshuai Zhao, Shuguang Liu, Dong Zhong, Boris Pinzur, Jie Zhang, Yang Wang, Jithin Jose, Hossein Pourreza, Jeff Baxter, Kushal Datta, Prabhat Ram, Luke Melton, and Joe Chau, Microsoft; Peng Cheng, Yongqiang Xiong, and Lidong Zhou, Microsoft Research **Removing Obstacles before Breaking Through the Memory Wall: A Close Look at HBM Errors in the Field** Ronglong Wu, Shuyue Zhou, Jiahao Lu, Zhirong Shen, Yiming Zhang, and Zikang Xu, Xiamen University; Kunlin Yang and Feilong Lin, Huawei Technologies Co., Ltd **MSFRD: Mutation Similarity based SSD Failure Rating and Diagnosis for Complex and Volatile Production Environments** Yuqi Zhang, Tianyi Zhang, Wenwen Hao, Shuyang Wang, Na Liu, and Xing He, Samsung RandD Institute China Xi'an, Samsung Electronics; Yang Zhang, Weixin Wang, Yongguang Cheng, Huan Wang, Jie Xu, Feng Wang, and Bo Jiang, ByteDance Inc.; Yongwong Gwon, Jongsung Na, Zoe Kim, and Geunrok Oh, Samsung Electronics
6:00 pm–7:30 pm	USENIX ATC '24 Poster Session and Reception		

Fri, July 12	OSDI '24	USENIX ATC '24	USENIX ATC '24
8:00 am–9:00 am	Continental Breakfast		
9:00 am–10:20 am	Cloud Security **DSig: Breaking the Barrier of Signatures in Data Centers** Marcos K. Aguilera, VMware Research; Clément Burgelin, Rachid Guerraoui, and Antoine Murat, EPFL; Athanasios Xygkis, Oracle Labs; Igor Zablotchi, Mysten Labs **Ransom Access Memories: Achieving Practical Ransomware Protection in Cloud with DeftPunk** 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 Inc. **Secret Key Recovery in a Global-Scale End-to-End Encryption System** Graeme Connell, Signal Messenger; Vivian Fang, UC Berkeley; Rolfe Schmidt, Signal Messenger; Emma Dauterman and Raluca Popa, UC Berkeley **Flock: A Framework for Deploying On-Demand Distributed Trust** Darya Kaviani and Sijun Tan, UC Berkeley; Pravein Govindan Kannan, IBM Research; Raluca Ada Popa, UC Berkeley	Deployed Systems **Diagnosing Application-network Anomalies for Millions of IPs in Production Clouds** Zhe Wang, Shanghai Jiao Tong University, China; Huanwu Hu, Alibaba Group, China; Linghe Kong, Shanghai Jiao Tong University, China; Xinlei Kang and Teng Ma, Alibaba Group, China; Qiao Xiang, Xiamen University, China; Jingxuan Li and Yang Lu, Alibaba Group, China; Zhuo Song, Alibaba Group and Shanghai Jiao Tong University, China; Peihao Yang, Alibaba Group, China; Jiejian Wu, Shanghai Jiao Tong University, China; Yong Yang and Tao Ma, Alibaba Group, China; Zheng Liu, Alibaba Group and Zhejiang University, China; Xianlong Zeng and Dennis Cai, Alibaba Group, China; Guihai Chen, Shanghai Jiao Tong University, China **Data Caching for Enterprise-Grade Petabyte-Scale OLAP** Chunxu Tang and Bin Fan, Alluxio; Jing Zhao and Chen Liang, Uber; Hope Wang and Beinan Wang, Alluxio; Ziyue Qiu, Carnegie Mellon University; Lu Qiu, Bowen Ding, Shouzhuo Sun, Saiguang Che, Jiaming Mai, Shouwei Chen, Yu Zhu, and Jianjian Xie, Alluxio; Yutian Sun, Meta; Yao Li and Yangjun Zhang, Uber; Ke Wang, Meta **Full Lifecycle Data Analysis on a Large-scale and Leadership Supercomputer: What Can We Learn from It?** Bin Yang, Tsinghua University, National Supercomputer Center in Wuxi; Hao Wei, Tsinghua University; Wenhao Zhu, Shandong University, National Supercomputer Center in Wuxi; Yuhao Zhang, Tsinghua University; Weiguo Liu, Shandong University; Wei Xue, Tsinghua University	Wide Area Network **Panorama: Optimizing Internet-scale Users' Routes from End to End** Geng Li, Shuihai Hu, and Kun Tan, Huawei **Enhancing Resource Management of the World's Largest PCDN System for On-Demand Video Streaming** Rui-Xiao Zhang, University of Illinois Urbana-Champaign; Haiping Wang, Shu Shi, Xiaofei Pang, Yajie Peng, and Zhichen Xue, ByteDance; Jiangchuan Liu, Simon Fraser University **TileClipper: Lightweight Selection of Regions of Interest from Videos for Traffic Surveillance** Shubham Chaudhary and Aryan Taneja, IIIT Delhi, India; Anjali Singh, IGDTUW Delhi, India; Purbasha Roy, Sohum Sikdar, Mukulika Maity, and Arani Bhattacharya, IIIT Delhi, India
10:20 am–10:50 am	Break with Refreshments		
10:50 am–12:10 pm	Data Management **FairyWREN: A Sustainable Cache for Emerging Write-Read-Erase Flash Interfaces** Sara McAllister and Yucong Wang, Carnegie Mellon University; Benjamin Berg, UNC Chapel Hill; Daniel S. Berger, Microsoft Azure and UW; George Amvrosiadis, Nathan Beckmann, and Greg Ganger, Carnegie Mellon University **Massively Parallel Multi-Versioned Transaction Processing** Shujian Qian and Ashvin Goel, University of Toronto **Burstable Cloud Block Storage with Data Processing Units** Junyi Shu, Peking University and Alibaba Group; Kun Qian and Ennan Zhai, Alibaba Group; Xuanzhe Liu and Xin Jin, Peking University **Motor: Enabling Multi-Versioning for Distributed Transactions on Disaggregated Memory** Ming Zhang, Yu Hua, and Zhijun Yang, Huazhong University of Science and Technology	Virtualization **Expeditious High-Concurrency MicroVM SnapStart in Persistent Memory with an Augmented Hypervisor** Xingguo Pang, Yanze Zhang, Liu Liu, and Xiaobo Zhou, University of Macau; Dazhao Cheng, WuHan University; Chengzhong Xu, University of Macau **Taming Hot Bloat Under Virtualization with HugeScope** Chuandong Li, Peking University; Sai Sha, Beijing Huawei Digital Technologies; Diyu Zhou, École Polytechnique Fédérale de Lausanne (EPFL); Yangqing Zeng, Xiran Yang, Yingwei Luo, and Xiaolin Wang, Peking University; Zhenlin Wang, Michigan Tech **CrossMapping: Harmonizing Memory Consistency in Cross-ISA Binary Translation** Chen Gao and Xiangwei Meng, Lanzhou University; Wei Li, Tsinghua University; Jinhui Lai, Lanzhou University; Yiran Zhang, Beijing University of Posts and Telecommunications; Fengyuan Ren, Lanzhou University and Tsinghua University	Security 2 **Efficient Decentralized Federated Singular Vector Decomposition** Di Chai, Junxue Zhang, and Liu Yang, Hong Kong University of Science and Technology; Yilun Jin, The Hong Kong University of Science and Technology; Leye Wang, Peking University; Kai Chen and Qiang Yang, Hong Kong University of Science and Technology **Models on the Move: Towards Feasible Embedded AI for Intrusion Detection on Vehicular CAN Bus** He Xu, Di Wu, and Yufeng Lu, Hunan University; Haibo Zeng, Virginia Tech; Jiwu Lu, Hunan University **(MY TALK)** **CPC: Flexible, Secure, and Efficient CVM Maintenance with Confidential Procedure Calls** Jiahao Chen, Zeyu Mi, Yubin Xia, Haibing Guan, and Haibo Chen, Shanghai Jiao Tong University
12:10 pm–1:40 pm	Lunch (on your own)		
1:40 pm–3:20 pm	Analysis of Correctness **Detecting Logic Bugs in Database Engines via Equivalent Expression Transformation** Zu-Ming Jiang and Zhendong Su, ETH Zurich **Inductive Invariants That Spark Joy: Using Invariant Taxonomies to Streamline Distributed Systems Proofs** Tony Nuda Zhang, University of Michigan — Ann Arbor; Travis Hance, Carnegie Mellon University; Manos Kapritsos, University of Michigan — Ann Arbor; Tej Chajed, UW-Madison; Bryan Parno, Carnegie Mellon University **Performance Interfaces for Hardware Accelerators** Jiacheng Ma, Rishabh Iyer, Sahand Kashani, Mahyar Emami, Thomas Bourgeat, and George Candea, EPFL **IronSpec: Increasing the Reliability of Formal Specifications** Eli Goldweber, Weixin Yu, Seyed Armin Vakil Ghahani, and Manos Kapritsos, University of Michigan **Identifying On-/Off-CPU Bottlenecks Together with Blocked Samples** Minwoo Ahn and Jeongmin Han, Sungkyunkwan University; Youngjin Kwon, Korea Advanced Institute of Science and Technology (KAIST); Jinkyu Jeong, Yonsei University	Storage 2 **RL-Watchdog: A Fast and Predictable SSD Liveness Watchdog on Storage Systems** Jinyong Ha, Seoul National University; Sangjin Lee, Chung-Ang University; Heon Young Yeom, Seoul National University; Yongseok Son, Chung-Ang University **Exploit both SMART Attributes and NAND Flash Wear Characteristics to Effectively Forecast SSD-based Storage Failures in Clusters** Yunfei Gu and Chentao Wu, Shanghai Jiao Tong University; Xubin He, Temple University **StreamCache: Revisiting Page Cache for File Scanning on Fast Storage Devices** Zhiyue Li and Guangyan Zhang, Tsinghua University **Scalable Billion-point Approximate Nearest Neighbor Search Using SmartSSDs** Bing Tian, Haikun Liu, Zhuohui Duan, Xiaofei Liao, and Hai Jin, Huazhong University of Science and Technology; Yu Zhang, Service Computing Technology and System Lab, Huazhong University of Science and Technology	Hardware **gVulkan: Scalable GPU Pooling for Pixel-Grained Rendering in Ray Tracing** Yicheng Gu, Yun Wang, Yunfan Sun, Yuxin Xiang, Yufan Jiang, Xuyan Hu, Zhengwei Qi, and Haibing Guan, Shanghai Jiao Tong University **vFPIO: A Virtual I/O Abstraction for FPGA-accelerated I/O Devices** Jiyang Chen, Harshavardhan Unnibhavi, Atsushi Koshiba, and Pramod Bhatotia TU Munich; **ScalaCache: Scalable User-Space Page Cache Management with Software-Hardware Coordination** Li Peng and Yuda An, Peking University; You Zhou, Huazhong University of Science and Technology; Chenxi Wang, Chinese Academy of Sciences; Qiao Li, Xiamen University; Cheng Chuanning, Huawei; Jie Zhang, Peking University **Centimani: Enabling Fast AI Accelerator Selection for DNN Training with a Novel Performance Predictor** Zhen Xie, Binghamton University; Murali Emani, Argonne National Laboratory; Xiaodong Yu, Stevens Institute of Technology; Dingwen Tao, Indiana University; Xin He, Guangzhou Institute of Technology, Xidian University; Pengfei Su, University of California, Merced; Keren Zhou, George Mason University; Venkatram Vishwanath, Argonne National Laboratory
3:20 pm–3:40 pm	Break with Refreshments		
3:40 pm–5:20 pm	ML Scheduling **dLoRA: Dynamically Orchestrating Requests and Adapters for LoRA LLM Serving** Bingyang Wu, Ruidong Zhu, and Zili Zhang, Peking University; Peng Sun, Shanghai AI Lab; Xuanzhe Liu and Xin Jin, Peking University **Parrot: Efficient Serving of LLM-based Applications with Semantic Variable** Chaofan Lin, Shanghai Jiao Tong University and Microsoft Research; Zhenhua Han, Chengruidong Zhang, and Yuqing Yang, Microsoft Research; Fan Yang, Microsoft Research Asia; Chen Chen, Shanghai Jiao Tong University; Lili Qiu, UT Austin and MSR Asia Shanghai **USHER: Holistic Interference Avoidance for Resource Optimized ML Inference** Sudipta Saha Shubha and Haiying Shen, University of Virginia; Anand Iyer, Georgia Institute of Technology **Fairness in Serving Large Language Models** Ying Sheng, Stanford University; Shiyi Cao, Dacheng Li, Banghua Zhu, and Zhuohan Li, UC Berkeley; Danyang Zhuo, Duke University; Joseph Gonzalez and Ion Stoica, UC Berkeley **MonoInfer: Enabling a New Monolithic Optimization Space for Neural Network Inference Tasks on Modern GPU-Centric Architectures** Donglin Zhuang, The University of Sydney; Zhen ZHENG, Alibaba Group; Haojun Xia, University of Sydney; Xiafei Qiu, Junjie Bai, and Wei Lin, Alibaba Group; Shuaiwen Leon Song, Microsoft/University of Sydney	Potpourri **A Difference World: High-performance, NVM-invariant, Software-only Intermittent Computation** Harrison Williams, Saim Ahmad, and Matthew Hicks, Virginia Tech **Efficient Large Graph Processing with Chunk-Based Graph Representation Model** Rui Wang, Weixu Zong, Shuibing He, Xinyu Chen, Zhenxin Li, and Zheng Dang, Zhejiang University **SlimArchive: A Lightweight Architecture for Ethereum Archive Nodes** Hang Feng, Yufeng Hu, and Yinghan Kou, Zhejiang University; Runhuai Li and Jianfeng Zhu, BlockSec; Lei Wu and Yajin Zhou, Zhejiang University **Every Mapping Counts in Large Amounts: Folio Accounting** David Hildenbrand and Martin Schulz, Technical University of Munich; Nadav Amit, Technion, Israel Institute of Technology 5:10 end
5:20 pm–5:30 pm	**Closing Remarks** **Program Co-Chairs: Ada Gavrilovska, Georgia Institute of Technology; Douglas B. Terry, Amazon Web Services**	**Closing Remarks** **Saurabh Bagchi, Purdue University; Yiying Zhang, University of California, San Diego**	