Lianjie Cao

+1 (650) 236-7070 lianjie.cao@hpe.com Website Google Scholar LinkedIn
820 N McCarthy Blvd, Milpitas, CA 95035

RESEARCH INTERESTS

My research focuses on optimizing system performance and resource scheduling across different layers (*e.g.*, application, system, and hardware) and different computer systems (*e.g.*, NFV, machine learning systems, serverless, 5G, and storage systems).

EDUCATION

Purdue University, West Lafayette, Indiana, USA

- Ph.D. in Computer Science — Aug 2011 – May 2018
  - Thesis: Data-driven Resource Allocation in Virtualized Environments
  - Adviser: Sonia Fahmy

George Mason University, Fairfax, Virginia, USA

- M.S. in Computer Engineering — Aug 2008 – Aug 2011
  - Thesis: A Rate-based Congestion Control Overlay System
  - Advisor: Brian Mark

Huazhong University of Science & Technology, Wuhan, Hubei, China

- B.E. in Artificial Intelligence and Automation — Sep 2004 – Jun 2008
  - Thesis: Email Monitoring System in WLAN
  - Advisor: Xiaoya Hu

PROFESSIONAL EXPERIENCE

HPE Labs

Senior Research Scientist — Feb 2023 – Present
Research Scientist — Aug 2018 – Jan 2023
Ph.D. Research Associate Intern — May 2014 – May 2017

Hewlett Packard

Software Development Intern — May 2013 – Jan 2014

Purdue University

Research Assistant — Aug 2011 – May 2018

PUBLICATIONS

Conference

[23] Weishu Deng, Yujie Yang, Peiran Du, Lingfeng Xiang, Zhen Lin, Chen Zhong, Faraz Ahmed, Lianjie Cao, Puneet Sharma, Song Jiang, Hui Lu, Jia Rao, "Scaling Attention Beyond GPUs for LLM Inference," In Proceedings of the 35th ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC), July 2026.

[22] Suyeon Lee, Khaled Diab, Diman Zad Tootaghaj, Lianjie Cao, Puneet Sharma, Ada Gavrilovska, "Griffin: Coherency-Aware Task Scheduling and Memory Allocation for CXL Interconnects," In Proceedings of the 40th ACM International Conference on Supercomputing (ICS), July 2026.

[21] Yi Li, Lianjie Cao, Faraz Ahmed, Puneet Sharma, Bingzhe Li, "Hippocampus: An Efficient and Scalable Memory Module for Agentic AI," In Proceedings of the 9th Machine Learning and Systems (MLSys), May 2026.

[20] Maximilian Pazer, Hardik Soni, Faraz Ahmed, Lianjie Cao, Khaled Diab, Noah Clemons, Ryan Hankins, Ed Benson, Bruck Girmay, Puneet Sharma, "Doppler: LLM-Powered Comparative Analysis for Network Performance", In Proceedings of the Cray User Group (CUG), April 2026.

[19] Diman Zad Tootaghaj, Anna Yue, Lianjie Cao, Bob Lantz, Torsten Wilde, Bryan P Murray, Gourav Rattihalli, Abhishek Acharya, David Sydow, Puneet Sharma, "HPM4AI: Adapting HPM for Energy-Efficient LLM Inference," In Proceedings of the Cray User Group (CUG), April 2026.

[18] Umakant Sunil Kulkarni, Khaled Diab, Lianjie Cao, Faraz Ahmed, Shivang Aggarwal, Sonia Fahmy, Puneet Sharma, "Maestro: QoE-Aware Dynamic Resource Allocation in Wi-Fi Networks," In Proceedings of the 21st ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT), 12 pp., December 2025.

[17] Zhen Lin, Yujie Yang, Lingfeng Xiang, Lianjie Cao, Faraz Ahmed, Jia Rao, Hui Lu, Puneet Sharma, "Can Hardware Outsmart Software in Tiered Memory Management? A CMM-H Case Study," In Proceedings of the 18th ACM International Systems and Storage Conference (SYSTOR), 5 pp., September 2025.

[16] Jinsun Yoo, ChonLam Lao, Lianjie Cao, Bob Lantz, Minlan Yu, Tushar Krishna, Puneet Sharma, "Towards an Easy and Realistic Network Infrastructure Testing in Largescale Machine Learning," In Proceedings of the 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI), Poster, April 2025.

[15] Yunming Xiao, Diman Zad Tootaghaj, Aditya Dhakal, Lianjie Cao, Puneet Sharma, Aleksandar Kuzmanovic, "Conspirator: SmartNIC-Aided Control Plane for Distributed ML Workloads," In Proceedings of the USENIX Annual Technical Conference (USENIX ATC), pp. 767–784, July 2024.

[14] Ali Tariq, Lianjie Cao, Faraz Ahmed, Eric Rozner, Puneet Sharma, "Accelerating Containerized Machine Learning Workloads," In Proceedings of the IEEE Network Operations and Management Symposium (NOMS), 10 pp., May 2024.

[13] Umakant Kulkarni, Khaled Diab, Shivang Aggarwal, Lianjie Cao, Faraz Ahmed , Puneet Sharma, Sonia Fahmy, "Understanding the Impact of Wi-Fi Configuration on Volumetric Video Streaming Applications," In Proceedings of the 15th ACM SIGCOMM Workshop on Emerging Multimedia Systems (EMS), 6 pp., September 2023.

[12] Zhen Lin, Lianjie Cao, Faraz Ahmed, Hui Lu, Puneet Sharma, "When Caching Systems Meet Emerging Storage Devices: A Case Study," In Proceedings of the 15th ACM Workshop on Hot Topics in Storage and File Systems (HotStorage), 6 pp., July 2023.

[11] Amit Samanta, Faraz Ahmed, Lianjie Cao, Ryan Stutsman, Puneet Sharma, "Persistent Memory-Aware Scheduling for Serverless Workloads," In Proceedings of the 4th Workshop on Extreme-Scale Storage and Analysis (ESSA) co-located with IEEE IPDPS, 6 pp., May 2023.

[10] Junguk Cho, Diman Zad Tootaghaj, Lianjie Cao, Puneet Sharma, "SLA-Driven ML Inference Framework For Clouds With Heterogeneous Accelerators," In Proceedings of the 5th Machine Learning and Systems (MLSys), pp. 20–32, August 2022.

[9] Lianjie Cao, Puneet Sharma, "Co-locating Containerized Workload Using Service Mesh Telemetry," In Proceedings of the 17th ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT), pp. 168–174, December 2021.

[8] Lianjie Cao, Anu Mercian, Diman Zad Tootaghaj, Faraz Ahmed, Puneet Sharma, Vinay Saxena, "eCaaS: A Management Framework of Edge Container as a Service for Business Workload," In Proceedings of the 4th ACM International Workshop on Edge Systems, Analytics and Networking (EdgeSys) co-located with EuroSys, pp. 73–78, April 2021.

[7] Lianjie Cao, Sonia Fahmy, Puneet Sharma, "Data-driven Resource Allocation in Virtualized Environments," In Proceedings of the 16th IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 659–664, April 2019. Dissertation Paper.

[6] Lianjie Cao, Sonia Fahmy, Puneet Sharma, Shandian Zhe, "Data-driven Resource Flexing for Network Functions Virtualization," In Proceedings of the 14th ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), pp. 111–124, July 2018.

[5] Lianjie Cao, Xiangyu Bu, Sonia Fahmy, Siyuan Cao, "Towards High Fidelity Network Emulation," In Proceedings of the 26th IEEE International Conference on Computer Communications and Networks (ICCCN), 11 pp., July 2017. (Invited)

[4] Lianjie Cao, Puneet Sharma, Sonia Fahmy, Vinay Saxena, "ENVI: Elastic resource flexing for Network function VIrtualization," In Proceedings of the 9th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud), 6 pp., July 2017.

[3] Amit Sheoran, Xiangyu Bu, Lianjie Cao, Puneet Sharma, and Sonia Fahmy, "An Empirical Case for Container- driven Fine-grained VNF Resource Flexing," In Proceedings of the 2nd IEEE Conference on Network Function Virtualization & Software Defined Networks (NFV-SDN), pp. 121–127, November 2016. **Best Paper Runner-up**

[2] Lianjie Cao, Puneet Sharma, Sonia Fahmy, and Vinay Saxena, "NFV-VITAL: A Framework for Characterizing the Performance of Virtual Network Functions," In Proceedings of the 1st IEEE Conference on Network Function Virtualization & Software Defined Networks (NFV-SDN), pp. 93–99, November 2015. **Best Paper Award**

[1] Lianjie Cao, Thibaut Provost, and Ramana Kompella, “PhishLive: A View of Phishing and Malware Attacks from an Edge Router,” In Proceedings of the 14th International Conference on Passive and Active Measurement (PAM), pp. 239–249, March 2013.

Journal

[4] Choi Sean, Patel Disha, Zad Tootaghaj Diman, Cao Lianjie, Ahmed Faraz, Sharma Puneet, “FedNIC: Enhancing Privacy-Preserving Federated Learning via Homomorphic Encryption Offload on SmartNIC,” In Frontiers in Computer Science, vol. 6, 15 pp., October 2024.

[3] Faraz Ahmed, Lianjie Cao, Ayush Goel, Puneet Sharma, “NeIL: Intelligent Replica Selection for Distributed Applications,” In IEEE Transactions on Machine Learning in Communications and Networking (TMLCN), vol. 2, pp. 1580–1594, October 2024.

[2] Arun Raghuramu, Lianjie Cao, Puneet Sharma, Mario Sanchez, Joon-Myung Kang, Chen-Nee Chuah, David Lee, Vinay Saxena, “Metered Boot: Trusted Framework for Application Usage Rights Management in Virtualized Ecosystems,” In IEEE Transactions on Network and Service Management (TNSM), vol. 19, no. 3, pp. 2238–2250, September 2022.

[1] Amit Sheoran, Sonia Fahmy, Lianjie Cao, Puneet Sharma, “AI-Driven Provisioning in the 5G Core,” In IEEE Internet Computing, vol. 25, no. 2, pp. 18–25, 1 March-April 2021.

Patent

[30] Shivang Aggarwal, Karthik Murthy, Khaled Diab, Selva Ulaganathan Jeyakumar, Lianjie Cao, Rajarshi Bhattacharyya, Faraz Ahmed, Sachin Ganu, Puneet Sharma, Intelligent and Efficient Flow Scheduling for Wi-Fi Networks. Approved for filing.

[29] Khaled Diab, Lianjie Cao,, Diman Zad Tootaghaj, Puneet Sharma, Sumanth Umesh, Cache Snoop Filters In Network Switches. US Patent Application Filed.

[28] Lianjie Cao, Divya Kiran Kadiyala, Puneet Sharma, Managing Shared Memory for Artificial Intelligence Model Training. US Patent Application Filed.

[27] Lianjie Cao, Yi Li, Faraz Ahmed, Puneet Sharma, Memory Architecture for Contextual Data Retrieval for Agentic Artificial Intelligence. US Patent Application Filed.

[26] Khaled Diab, Lianjie Cao, Puneet Sharma, Weigao Su, Scalable Multipath GPU Communication. US Patent Application Filed.

[25] Ayush Goel, Khaled Diab, Faraz Ahmed, Diman Zad Tootaghaj, Lianjie Cao,, Hardik Soni, Puneet Sharma, Improved Serverless LLM Loading. US Patent Application Filed.

[24] Diman Zad Tootaghaj, Lianjie Cao, Puneet Sharma, Job Scheduling Based on Carbon Emission Information and A Power Consumption Reduction Objective. US Patent Application Filed.

[23] Khaled Diab, Suyeon Lee, Diman Zad Tootaghaj, Lianjie Cao, Puneet Sharma, Joint Task Scheduling and Memory Allocation in Cache-coherent Memory Interconnects. US Patent Application Filed.

[22] Lianjie Cao, Faraz Ahmed, Hana Khamfroush, Puneet Sharma, “Determining Resource Allocations for Microservice-based Applications,” US Patent Application 18/894,478.

[21] Diman Zad Tootaghaj, Yunming Xiao, Aditya Dhakal, Puneet Sharma, Lianjie Cao, “Job Allocations To Fractions of Parallel Processing Units (PPUs),” US Patent Application 18/765,440.

[20] Diman Zad Tootaghaj, Yunming Xiao, Aditya Dhakal, Puneet Sharma, Lianjie Cao, “DMA Transfers of Job Data From An Adapter To Parallel Processing Unit (PPU) Fractions,” US Patent Application 18/765,445.

[19] Lianjie Cao, Saeed Rashidi, Puneet Sharma, Garrett Goon, Paolo Faraboschi, “Resilient Optimizer States for Fully Sharded Data Parallel,” US Patent Application 18/648,217.

[18] Lianjie Cao, Saeed Rashidi, Puneet Sharma, Garrett Goon, Paolo Faraboschi, “Resilient Fully Sharded Data Parallel,” US Patent Application 18/634,065.

[17] Saeed Rashidi, Lianjie Cao, Puneet Sharma, Khaled Diab, Hardik Soni, Faraz Ahmed, “Multi-tenant Collective Communication Fabric,” US Patent Application 18/538,936.

[16] Lianjie Cao, Zhen Lin, Faraz Ahmed, Puneet Sharma, “Accelerating Containerized Applications with Caching,” US Patent 12,367,152 B2.

[15] Chinlin Chen, Uyen Chau, Faraz Ahmed, Lianjie Cao, “Dynamic Device Persona Identification in a Network,” US Patent 12,407,567 B2.

[14] Faraz Ahmed, Lianjie Cao, Puneet Sharma, "Optimizing Cost and Performance for Serverless Data Analytics Workloads," US Patent Application 18/175,411.

[13] Faraz Ahmed, Lianjie Cao, Puneet Sharma, "Enabling Persistent Memory for Serverless Applications," US Patent Application 17/957,700.

[12] Faraz Ahmed, Lianjie Cao, Puneet Sharma, "Machine Learning-based Approaches for Service Function Chain Selection," US Patent 12,133,095 B2.

[11] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Ali Tariq, "Network-aware Resource Allocation - II," US Patent 12,132,668 B2.

[10] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Ali Tariq, "Network-aware Resource Allocation - I," US Patent 11,665,106 B2.

[9] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Anu Mercian, Diman Zad Tootaghaj, "Deployment and Configuration of an Edge Site Based on Declarative Intents Indicative of a Use Case - II," US Patent 11,914,982 B2.

[8] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Anu Mercian, Diman Zad Tootaghaj, "Deployment and Configuration of an Edge Site Based on Declarative Intents Indicative of a Use Case - I," US Patent 11,698,780 B2.

[7] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Ali Tariq, "Systems and Methods of Resource Configuration Optimization for Machine Learning Workloads – IV," US Patent Application 18/654,953.

[6] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Ali Tariq, "Systems and Methods of Resource Configuration Optimization for Machine Learning Workloads – III," US Patent 12,141,608 B2.

[5] Lianjie Cao, Puneet Sharma, Faraz Ahmed, Ali Tariq, "Systems and Methods of Resource Configuration Optimization for Machine Learning Workloads – II," US Patent 12,001,511 B2.

[4] Lianjie Cao, Puneet Sharma, Faraz Ahmed, "Systems and Methods of Resource Configuration Optimization for Machine Learning Workloads – I," US Patent 11,797,340 B2.

[3] Lianjie Cao, Puneet Sharma, "Assignment of Microservices," US Patent 10,827,020 B1.

[2] Lianjie Cao, Puneet Sharma, Vinay Saxena, "Virtual Network Function Resource Allocation," US Patent 11,010,205 B2.

[1] Lianjie Cao, Puneet Sharma, Vinay Saxena, Vasu Sankhavaram, Badrinath Natarajan, "Determining Virtual Network Function Configurations," US Patent 10,489,180 B2.

PROFESSIONAL ACTIVITIES

Technical Program Committee

- Chair, IEEE Cloud Summit 2026 - Industry Track
- IEEE/IFIP Network Operations and Management Symposium (NOMS) 2022, 2023, 2024, 2025, 2026
- IEEE International Conference on Network Softwarization (NetSoft) - PhD symposium 2023, 2025
- IEEE/ACM International Symposium on Quality of Service (IWQoS) 2023
- Passive and Active Measurement (PAM) 2022
- IFIP/IEEE International Symposium on Integrated Network Management (IM) 2021
- IFIP Networking 2019, 2020

Invited Reviewer

- IEEE Computer Architecture Letters
- IEEE Transactions on Network and Service Management (TNSM)
- IEEE Transactions on Network Science and Engineering (TNSE)
- IEEE Sensors Journal
- IEEE Networking Letters
- IEEE International Conference on Network Softwarization (NetSoft)

External Reviewer

- ACM SIGCOMM Symposium on SDN Research (SOSR)
- ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT)
- IEEE International Conference on Computer Communications (INFOCOM)

SKILLS

Programming Languages

Python, Bash, C/C++, Golang, Java, Matlab, R, LaTeX, JavaScript, HTML, SQL

Tools
TensorFlow, PyTorch, Kubernetes, Docker, OpenStack, Open vSwtich, Mininet

[*Updated on 2026-05-19*]