Yue Cheng

Associate Professor of Data Science and Computer Science
University of Virginia

31 Bonnycastle Dr
Charlottesville, VA 22093
✉ *mrz7dp@virginia.edu*
tddg.github.io

Research Interests

Storage systems, data systems, cloud computing, serverless computing, high-performance computing, operating systems, systems for ML, ML for systems
The overarching goal of my research is to enable practical, efficient, and easy-to-use computer systems for the growing data demands of modern high-end applications running on existing as well as emerging computing platforms. My current research focuses on: (1) designing efficient serverless computing and data systems using a full-stack approach spanning application frameworks, platforms, operating systems, and hardware; (2) building scalable and efficient data systems and ML systems; and (3) utilizing ML approaches to advance the computing and storage systems.

Professional Experience and Employment

08/2023–present **Associate Professor**, *University of Virginia*, Charlottesville, VA.
School of Data Science and SEAS Department of Computer Science

08/2022–08/2023 **Assistant Professor**, *University of Virginia*, Charlottesville, VA.
School of Data Science and SEAS Department of Computer Science

08/2017–08/2022 **Assistant Professor**, *George Mason University*, Fairfax, VA.
Department of Computer Science

2011–2017 **Research/Teaching Assistant**, *Virginia Tech*, Blacksburg, VA.
Department of Computer Science

06/2015–12/2015 **Research Intern**, *EMC*, Princeton, NJ.
Offline flash caching

05/2014–08/2014 **Research Intern**, *IBM Research–Almaden*, San Jose, CA.
Cloud analytics storage tiering

05/2013–08/2013 **Research Intern**, *IBM Research–Almaden*, San Jose, CA.
Load balanced in-memory caching

Education

2011–2017 **Virginia Polytechnic Institute and State University (Virginia Tech)**, *Blacksburg, VA*.
Ph.D. in Computer Science

2005–2009 **Beijing University of Posts and Telecommunications (BUPT)**, *Beijing, China*.
B.Eng. in Computer Science

Awards & Honors

2023 **Outstanding Researcher Award**, for achievements in research at the University of Virginia

2023 **Samsung Global Research Outreach Award**, Samsung Advanced Institute of Technology and Samsung Memory Solutions Lab

2022 **IEEE CS TCHPC Early Career Researchers Award for Excellence in High Performance Computing**

2022 **Meta Research Award** of the Meta AI System Hardware/Software Codesign Competition

2022 **Best Student Paper Award Finalist** of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC 2022): 5 out of 81 accepted papers

2022 **Outstanding Teacher Award** of the Computer Science Department at George Mason University

2022 **Award Finalist** of Facebook (Meta) Mathematical Modeling & Optimization for Large-Scale Distributed Systems Award Competition

2021 **NSF CAREER Award** for the project titled "CAREER: Harnessing Serverless Functions to Build Highly Elastic Cloud Storage Infrastructure"

2020 **Amazon Research Award** for the project titled "Distributed Large-scale Graph Deep Learning by Gradient-free Optimization"

2012–2015 **Student Travel Grant:** USENIX ATC'15, ACM HPDC'15, EuroSys'15, USENIX OSDI'14, USENIX FAST'14, ACM SoCC'13, USENIX OSDI'12

2014 **Pratt Fellowship** (**Best Teaching Assistant Award**) awarded by Computer Science at Virginia Tech

2006–2009 **University Scholarship** awarded by Beijing University of Posts and Telecommunications, China

Publication

A: Students for whom I serve as the advisor; M: Students I mentor.

Manuscripts Currently Under Review/Revision

1. **A Closer Look into IPFS: Accessibility, Content, and Performance**.
   ACM SIGMETRICS / IFIP Performance 2024 (***SIGMETRICS'24***), (*One-shot revision*).
   Ruizhe Shi[M], Ruizhi Cheng, Bo Han, **Yue Cheng**, Songqing Chen.
2. **Everything You Always Wanted to Know About Storage Compressibility of Pre-Trained ML Models but Were Afraid to Ask**.
   50^{th} International Conference on Very Large Data Bases (***VLDB'24***), (*Under revision*).
   Zhaoyuan Su[A], Ammar Ahmed, Zirui Wang[A], Ali Anwar, **Yue Cheng**.
3. **Understanding the Readiness of Web 3.0: A Client's Perspective of IPFS**.
   The 2024 ACM Web Conference (***TheWebConf'24***), (*Under review*).
   Ruizhe Shi[M], Ruizhi Cheng, Yuqi Fu[A], Bo Han, **Yue Cheng**, Songqing Chen.
4. **Concurrency-Informed Orchestration for Serverless Functions**.
   ACM Special Interest Group on Data Communication 2024 (***SIGCOMM'24***), (*Under review*).
   Qichang Liu[M], **Yue Cheng**, Haiying Shen, Ao Wang[A], Bharathan Balaji.
5. **Staleness-Alleviated Distributed GNN Training via Online Dynamic-Embedding Prediction**.
   The 40^{th} IEEE International Conference on Data Engineering (***ICDE'24***), (*Under review*).
   Guangji Bai, Ziyang Yu, Zheng Chai[A], **Yue Cheng**, Liang Zhao.
6. **Distributed Graph Neural Network Training with Decoupled Representation Storage**.
   The 40^{th} IEEE International Conference on Data Engineering (***ICDE'24***), (*Under review*).
   Zheng Chai[A], Guangji Bai, Liang Zhao, **Yue Cheng**.
7. **FedCaSe: A Caching and Scheduling Framework for Federated Learning**.
   50^{th} International Conference on Very Large Data Bases (***VLDB'24***), (*Under review*).
   Redwan Ibne Seraj Khan[M], Arnab K. Paul, **Yue Cheng**, Ali R. Butt.
8. **Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models**.
   Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Carl Yang, **Yue Cheng**, Liang Zhao.
9. **SEALS: A Self-Adaptive, Learned Scheduler for Serverless Functions**.
   2024 USENIX Annual Technical Conference (***ATC'24***), (*Under review*).
   Yuqi Fu, Ruizhe Shi, Haoliang Wang, Songqing Chen, **Yue Cheng**.

Refereed Conference and Workshop Papers

★: **Top venues (18)**

Systems ASPLOS'23, FAST'23, FAST'20, FAST'18, ATC'21, ATC'16, SoCC'21, SoCC'20, EuroSys'15

HPC SC'22, SC'21, SC'18, HPDC'20, HPDC'16, HPDC'15

DB & ML VLDB'24 $\times 2$, VLDB'23

[VLDB '24]★ **Everything You Always Wanted to Know About Storage Compressibility of Pre-Trained ML Models but Were Afraid to Ask**.
50^{th} International Conference on Very Large Data Bases (***VLDB'24***), (*To appear*).
Zhaoyuan Su[A], Ammar Ahmed, Zirui Wang[A], Ali Anwar, **Yue Cheng**.

[VLDB '24]★ **Algorithmic Complexity Attacks for Dynamic Learned Indexes**.
50^{th} International Conference on Very Large Data Bases (***VLDB'24***), (*To appear*).
Rui Yang[A], Evgenios M. Kornaropoulos, **Yue Cheng**.

[BigData '23] **Towards Cost-effective and Resource-aware Aggregation at Edge for Federated Learning**.
2023 IEEE International Conference on Big Data (***BigData'23***), (AR: $92/526 = 17.5\%$).
Ahmad Khan, Yuze Li, Xinran Wang, Sabaat Haroon, Haider Ali, **Yue Cheng**, Ali R. Butt, and Ali Anwar.

[ASPLOS '23]★ **λFS: A Scalable and Elastic Distributed File System Metadata Service using Serverless Functions**.
ACM Conference on Architectural Support for Programming Languages and Operating Systems (AR: $50/238 = 21\%$), (*To appear*).
Benjamin Carver[A], Runzhou Han, Jingyuan Zhang[A], Mai Zheng, **Yue Cheng**.

[VLDB '23]★ **InfiniStore: Elastic Serverless Cloud Storage**.
49^{th} International Conference on Very Large Data Bases (***VLDB'23***).
Jingyuan Zhang[A], Ao Wang[A], Xiaolong Ma, Benjamin Carver[A], Nicholas John Newman[A], Ali Anwar, Vasily Tarasov, Lukas Rupprecht, Dimitrios Skourtis, Feng Yan, **Yue Cheng**.

[FAST '23]★ **SHADE: Enable Fundamental Cacheability for Distributed Deep Learning Training**.
USENIX Conference on File and Storage Techniques (***FAST'23***), (AR: $28/123 = 22.8\%$).
Redwan Ibne Seraj Khan[M], Ahmad Hossein Yazdani[M], Yuqi Fu[A], Arnab K. Paul, Bo Ji, Xun Jian, **Yue Cheng**, Ali R. Butt.

[SC '22]★ **SFS: Smarter OS Scheduling for Serverless Functions**.
The International Conference for High Performance Computing, Networking, Storage, and Analysis (***SC'22*** – ***Best Student Paper Award Finalist***), (AR: $81/320 = 25.3\%$).
Yuqi Fu[A], Li Liu[M], Haoliang Wang, **Yue Cheng**, Songqing Chen.

[SoCC '21]★ **Mind the Gap: Broken Promises of CPU Reservations in Containerized Multi-tenant Clouds**.
ACM Symposium on Cloud Computing (***SoCC'21***), (AR: $46/145 = 31.7\%$).
Li Liu[M], Haoliang Wang, An Wang, Mengbai Xiao, **Yue Cheng**, Songqing Chen.

[SC '21]★ **FedAT: A High-Performance and Communication-Efficient Federated Learning System with Asynchronous Tiers**.
The International Conference for High Performance Computing, Networking, Storage, and Analysis (***SC'21***), (AR: $86/365 = 23.6\%$).
Zheng Chai[A], Yujing Chen, Ali Anwar, Liang Zhao, **Yue Cheng**, Huzefa Rangwala.

[ATC '21]★ **FaaSNet: Scalable and Fast Provisioning of Custum Serverless Container Runtimes at Alibaba Cloud Function Compute**.
2021 USENIX Annual Technical Conference (***ATC'21***), (AR: $64/341 = 18.8\%$).
Ao Wang[A], Shuai Chang, Huangshi Tian, Hongqi Wang, Haoran Yang, Huiba Li, Rui Du, **Yue Cheng**.

[OPT '21] **Community-based Layerwise Distributed Training of Graph Convolutional Networks**.
NeurIPS 2021 Workshop on Optimization for Machine Learning (***OPT'21***).
Hongyi Li, Junxiang Wang, Yongchao Wang, **Yue Cheng**, Liang Zhao.

[ICDM '20] **Toward Model Parallelism for Deep Neural Network based on Gradient-free ADMM Framework**.
20^{th} IEEE International Conference on Data Mining (***ICDM'20***), (AR: $91/930 = 9.8\%$).
Junxiang Wang, Zheng Chai[A], **Yue Cheng**, Liang Zhao.

[SoCC '20]★ **Wukong: A Scalable and Locality-Enhanced Framework for Serverless Parallel Computing**.
ACM Symposium on Cloud Computing (***SoCC'20***), (AR: $35/143 = 24.5\%$).
Benjamin Carver[A], Jingyuan Zhang[A], Ao Wang[A], Ali Anwar, Panruo Wu, **Yue Cheng**.

[ICML WS '20] **Tunable Subnetwork Splitting for Model-parallelism of Neural Network Training**.
ICML 2020 Workshop on Beyond First-Order Methods in ML systems (***ICML WS'20***).
Junxiang Wang, Zheng Chai[A], **Yue Cheng**, Liang Zhao.

[HPDC '20]★ **TiFL: A Tier-based Federated Learning System**.
ACM Symposium on High-Performance Parallel and Distributed Computing (***HPDC'20***), (AR: $16/71 = 22.5\%$).
Zheng Chai[A], Ahsan Ali, Syed Zawad, Ali Anwar, Stacey Truex, Nathalie Baracaldo, Yi Zhou, Heiko Ludwig, Feng Yan, **Yue Cheng**.

[FAST '20]★ **InfiniCache: Exploiting Ephemeral Serverless Functions to Build a Cost-Effective Memory Cache**.
USENIX Conference on File and Storage Techniques (***FAST'20***), (AR: $23/138 = 16.7\%$).
Ao Wang[A] (co-primary), Jingyuan Zhang[A] (co-primary), Xiaolong Ma, Ali Anwar, Vasily Tarasov, Lukas Rupprecht, Dimitrios Skourtis, Feng Yan, **Yue Cheng**.

[PDSW '19] **In Search of a Fast and Efficient Serverless DAG Engine**.
The 4^{th} International Parallel Data Systems Workshop (***PDSW'19***).
Benjamin Carver[A], Jingyuan Zhang[A], Ao Wang[A], **Yue Cheng**.

[Cloud '19] **Bolt: Towards a Scalable Docker Registry**.
The IEEE International Conference on Cloud Computing (***Cloud'19***), (AR: 20.8%).
Michael Littley, Ali Anwar, Hannan Fayyaz[M], Zeshan Fayyaz[M], Vasily Tarasov, Lukas Rupprecht, Dimitrios Skourtis, Mohamed Mohamed, Heiko Ludwig, **Yue Cheng**, Ali R. Butt.

[OpML '19] **Towards Taming the Resource and Data Heterogeneity in Federated Learning**.
2019 USENIX Conference on Operational Machine Learning (***OpML'19***), (AR: $16/30 = 53.3\%$).
Zheng Chai[A], Hannan Fayyaz[M], Zeshan Fayyaz[M], Ali Anwar, Yi Zhou, Nathalie Baracaldo, Heiko Ludwig, **Yue Cheng**.

[VEE '19] **vCPU as a Container: Towards Accurate CPU Allocation for VMs**.
The 15^{th} ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (***VEE'19***), (AR: $15/33 = 45.5\%$).
Li Liu[M], Haoliang Wang, An Wang, Mengbai Xiao, **Yue Cheng**, Songqing Chen.

[BigData '18] **Analyzing Alibaba's Co-located Datacenter Workloads**.
IEEE International Conference on Big Data (***BigData'18***), (AR: 38.8%).
Yue Cheng, Ali Anwar, Xuejing Duan.

[SC '18]★ **BespoKV: Application Tailored Scale-Out Key-Value Stores**.
The International Conference for High Performance Computing, Networking, Storage, and Analysis (***SC'18***), (AR: $68/288 = 23.6\%$).
Ali Anwar, **Yue Cheng**, Hai Huang, Jingoo Han, Hyogi Sim, Dongyoon Lee, Fred Douglis, and Ali R. Butt.

[APSys '18] **Characterizing Co-located Datacenter Workloads: An Alibaba Case Study**.
The 9^{th} ACM SIGOPS Asia-Pacific Workshop on Systems (***APSys'18***), (AR: $21/50 = 42\%$).
Yue Cheng, Zheng Chai*, Ali Anwar.

[IPDPS '18] **Chameleon: An Adaptive Wear Balancer for Flash Clusters**.
IEEE International Parallel & Distributed Processing Symposium (***IPDPS'18***), (AR: $113/461 = 24.5\%$).
Nannan Zhao, Ali Anwar, **Yue Cheng**, Mohammed Salman, Daping Li, Jiguang Wan, Changsheng Xie, Xubin He, Feiyi Wang, and Ali R. Butt.

[FAST '18]★ **Improving Docker Registry Design based on Production Workload Analysis**.
USENIX Conference on File and Storage Techniques (***FAST'18***), (AR: $23/140 = 16.4\%$).
Ali Anwar, Mohamed Mohamed, Vasily Tarasov, Michael Littley, Lukas Rupprecht, **Yue Cheng**, Nannan Zhao, Dimitrios Skourtis, Amit S. Warke, Heiko Ludwig, Dean Hildebrand, Ali R. Butt.

[ATC '16]★ **Erasing Belady's Limitations: In Search of Flash Cache Offline Optimality**.
The 2016 USENIX Annual Technical Conference (***ATC'16***), (AR: $47/266 = 17.7\%$).
Yue Cheng, Fred Douglis, Philip Shilane, Michael Trachtman, Grant Wallace, Peter Desnoyers, and Kai Li.

[HotStorage '16] **ClusterOn: Building Highly Configurable and Reusable Clustered Data Services using Simple Data Nodes**.
The 8^{th} USENIX Workshop on Hot Topics in Storage and File Systems (***HotStorage'16***), (AR: $24/65 = 36.9\%$).
Ali Anwar, **Yue Cheng**, Hai Huang, and Ali R. Butt.

[HPDC '16]★ **MOS: Workload-aware Elasticity for Cloud Object Stores**.
The 25^{th} ACM Symposium on High-Performance Parallel and Distributed Computing (***HPDC'16***), (AR: $20/129 = 15.5\%$).
Ali Anwar, **Yue Cheng**, Aayush Gupta, and Ali R. Butt.

[VarSys '16] **Towards Managing Variability in the Cloud**.
The 1^{st} IEEE International Workshop on Variability in Parallel and Distributed Systems (***VarSys'16***).
Ali Anwar, **Yue Cheng**, and Ali R. Butt.

[PDSW '15] **Taming the Cloud Object Stores with MOS**.
The 10^{th} ACM Parallel Data Storage Workshop (***PDSW'15***), (AR: $9/25 = 36\%$).
Ali Anwar, **Yue Cheng**, Aayush Gupta, and Ali R. Butt.

[HotCloud '15] **Pricing Games for Hybrid Object Stores in the Cloud: Provider vs. Tenant**.
The 7^{th} USENIX Workshop on Hot Topics in Cloud Computing (***HotCloud'15***),(AR: $21/64 = 32.8\%$).
Yue Cheng, M. Safdar Iqbal, Aayush Gupta, and Ali R. Butt.

[HPDC '15]★ **Cast: Tiering Storage for Data Analytics in the Cloud**.
The 24^{th} ACM Symposium on High-Performance Parallel and Distributed Computing (***HPDC'15***), (AR: $19/116 = 16.4\%$).
Yue Cheng, M. Safdar Iqbal, Aayush Gupta, and Ali R. Butt.

[EuroSys '15]★ **An In-Memory Object Caching Framework with Adaptive Load Balancing**.
The 10^{th} ACM European Conference on Computer Systems (***EuroSys'15***), (AR: $32/154 = 20.8\%$).
Yue Cheng, Aayush Gupta, and Ali R. Butt.

Technical Reports

[VT technical report] **MOANA: Modeling and Analyzing I/O Variability in Parallel System Experimental Design**.
Kirk Cameron, Ali Anwar, **Yue Cheng**, Li Xu, Bo Li, Uday Ananth, Yili Hong, Layne T. Watson, and Ali R. Butt.

Posters and Demos

[NSDI '19] **HyperFaaS: A Truly Elastic Serverless Computing Framework**.
USENIX Symposium on Networked Systems Design and Implementation (***NSDI'19***), (Poster).
Jingyuan Zhang*, Ao Wang*, Min Li, Yuan Chen, **Yue Cheng**.

[APSys '15] **Taming the Cloud Object Stores with MOS**.
The 6^{th} ACM SIGOPS Asia-Pacific Workshop on Systems (***APSys'15***), (Poster).
Ali Anwar, **Yue Cheng**, Aayush Gupta, and Ali R. Butt.

[SoCC '13] **High Performance In-Memory Caching through Flexible Fine-Grained Services**.
2013 ACM Symposium on Cloud Computing (***SoCC'13***), (Poster).
Yue Cheng, Aayush Gupta, Anna Povzner, and Ali R. Butt.

Book Chapters

[Book chapter] **SDN helps Big Data to optimize storage**.
Big Data and Software Defined Networks, editor: Javid Taheri. IET, ISBN 978-1-78561-304-3. 2018.
Ali R. Butt, Ali Anwar, and **Yue Cheng**.

Refereed Journals

[TNNLS] **Community-based Distributed Training of Graph Convolutional Networks via ADMM**.
IEEE Transactions on Neural Networks and Learning Systems (***TNNLS***) (*Under review*).
Hongyi Li, Junxiang Wang, Yongchao Wang, **Yue Cheng**, Liang Zhao.

[TNNLS] **Towards Quantized Model Parallelism for Graph-Augmented MLPs Based on Gradient-Free ADMM Framework**.
IEEE Transactions on Neural Networks and Learning Systems (***TNNLS***).
Junxiang Wang, Hongyi Li, Zheng Chai[A], Yongchao Wang, **Yue Cheng**, Liang Zhao.

[TPDS] **Customizable Scale-Out Key-Value Stores**.
IEEE Transactions on Parallel and Distributed Systems (***TPDS***), Volume: 31, Issue: 9, Pages: 2081-2096, Apr. 25 2020, (Impact Factor $= 3.402$).
Ali Anwar, **Yue Cheng**, Hai Huang, Jingoo Han, Hyogi Sim, Dongyoon Lee, Fred Douglis, Ali R. Butt.

[TPDS] **MOANA: Modeling and Analyzing I/O Variability in Parallel System Experimental Design**.
IEEE Transactions on Parallel and Distributed Systems (***TPDS***), Volume: 30, Issue: 8, Pages: 1843-1856, Aug. 1 2019, (Impact Factor $= 3.402$).
Kirk Cameron, Ali Anwar, **Yue Cheng**, Li Xu, Bo Li, Uday Ananth, Yili Hong, Layne T. Watson, and Ali R. Butt.

[Internet Computing] **Provider versus Tenant Pricing Games for Hybrid Object Stores in the Cloud**.
IEEE Internet Computing's special issue on Cloud Storage: May/June 2016, Pages: 28-35, vol. 20.
Yue Cheng, M. Safdar Iqbal, Aayush Gupta, and Ali R. Butt.

Research Grants

14 awarded grants: 5 NSF grants + 8 industry awards/gifts (Adobe, Meta, Samsung, and Amazon) + 1 VA 4-initiatives project + 1 hardware donation.
Total grant amount: $4.1 M; Total personal share: $1.6 M.

Adobe Gift **"Serverless GPU and Storage Management for Large-scale, Interactive Machine Learning Training Workloads"**. Total: $25,000; My personal share: $25,000; Role: PI: Yue Cheng (UVA); Duration: 02/2024–present.

Samsung GRO **"Highly Efficient Pre-Trained LLM Storage with Near-Storage Compression and CXL Memory Integration"**. Total: $250,000; My personal share: $125,000; Role: PI: Yue Cheng (UVA), Co-PI: Ali Anwar (UMN); Duration: 01/2024–present.

Adobe Gift **"Serverless GPU and Storage Management for Large-scale, Interactive Machine Learning Training Workloads"**. Total: $20,000; My personal share: $20,000; Role: PI: Yue Cheng (UVA); Duration: 06/2023–present.

4-VA Collaborative Grant **"Near-Data Processing for Machine Learning Workloads Acceleration"**. Total: $35,000; My personal share: $5,000; Role: PI: Huaicheng Li (VT); Co-PI: Yue Cheng (UVA); Duration: 05/2023–present.

Meta Research Awards **"Serverless and Scalable GNN Training with Disaggregated Compute and Storage"**. Total: $50,000; My personal share: $25,000; Role: PI: Yue Cheng (UVA); Co-PI: Liang Zhao (Emory); Duration: 09/2022–08/2023.

Hardware **Western Digital Zoned Namespaces SSDs**. Two 4TB Western Digital ZN540 SSDs; Role: PI: Yue Cheng (UVA).

Adobe Gift **"Serverless GPU and Storage Management for Large-scale, Interactive Machine Learning Training Workloads"**. Total: $30,000; My personal share: $30,000; Role: PI: Yue Cheng (UVA); Duration: 05/2022–present.

Adobe Gift **"Serverless GPU and Storage Management for Large-scale, Interactive Machine Learning Training Workloads"**. Total: $10,000; My personal share: $10,000; PI: Yue Cheng (UVA); Duration: 09/2021–present.

NSF: CMMI-2134689 **"FMSG: Cyber: Federated Deep Learning for Future Ubiquitous Distributed Additive Manufacturing"**. Grant amount: $498,762; My personal share: $189,949 (38% share); PI: Jia Liu (Auburn); Co-PI: Yue Cheng (UVA); Duration: 10/01/2021–9/30/2023.

Adobe Gift **"Achieving Predictable Performance for FaaS Workloads via OS-Transparent Serverless Function Scheduling"**. Total: $10,000; My personal share: $10,000; PI: Yue Cheng (UVA); Duration: 03/2021–present

NSF: CNS-2045680 **"CAREER: Harnessing Serverless Functions to Build Highly Elastic Cloud Storage Infrastructure"**. Grant amount: $572,897 + $16,000 REU; My personal share: $572,897 + $16,000 REU (100% share); PI: Yue Cheng (UVA); Duration: 02/15/2021–02/14/2026.

Amazon Research Award **"Distributed Large-scale Graph Deep Learning by Gradient-free Optimization"**. Grant amount: $75,000; My personal share: $36,000; PI: Liang Zhao (Emory); Co-PI: Yue Cheng (UVA); Duration: 11/01/2020–10/31/2022.

NSF: MRI-2018631 **"MRI: Acquisition of an Adaptive Computing Infrastructure to Support Compute- and Data-Intensive Multidisciplinary Research"**. Grant amount: $750,000; PI: Elise Miller-Hooks (GMU); Co-PIs: Jayshree Sarma, Yue Cheng, Shobita Satyapal, Maria Emelianenko (GMU); Involved in designing Hopper, GMU's next-generation on-campus HPC Infrastructure; Duration: 08/01/2020–7/31/2023.

NSF: OAC-2007976 **"OAC Core: SMALL: DeepJIMU: Model-Parallelism Infrastructure for Large-scale Deep Learning by Gradient-Free Optimization"**. Grant amount: $498,609; My personal share: $249,302 (50% share); PI: Liang Zhao (Emory); Co-PI: Yue Cheng (UVA); Duration: 10/01/2020–9/30/2023.

NSF: CCF-1919075 **"SPX: Collaborative Research: Cross-stack Memory Optimizations for Boosting I/O Performance of Deep Learning HPC Applications"**. Grant amount: $1,273,487; UVA share: $320,603 (25% share); Role: PI: Yue Cheng (UVA); Duration: 10/01/2019–9/30/2023.

Time Allocation Grants

NSF CloudBank **"CAREER: Harnessing Serverless Functions to Build Highly Elastic Cloud Storage Infrastructure"**. Total: $33,230 AWS credit; PI: Yue Cheng (UVA); Duration: 07/21/2022–present.

IBM Cloud **"InfiniStore: Elastic Serverless Cloud Storage"**. Total: $24,000; PI: Yue Cheng (UVA); Duration: 12/30/2020–12/29/2021.

Google Cloud Platform **"Building a Purely Serverless Parallel Computing Framework"**. Total: $5,000; PI: Yue Cheng (UVA); Duration: 08/10/2020–08/09/2021.

Amazon Web Services **"LambDAG: A Lambda-aware DAG Engine"**. Total: $36,000; PI: Yue Cheng (UVA); Duration: 10/01/2019–10/31/2020.

Google Cloud Platform **"Building a Generic Serverless DAG Engine"**. Total: $10,000; PI: Yue Cheng (UVA); Duration: 08/20/2019–02/19/2020.

Google Cloud Platform **"Towards Serverless Computational Science"**. Total: $5,000; PI: Yue Cheng (UVA); Duration: 10/01/2018–07/31/2019.

Amazon Web Services **"Building a Virtual Serverless Cloud OS"**. Total: $36,000; PI: Yue Cheng (UVA); Duration: 08/01/2018–07/31/2019.

Talks

2023 **SHADE: Enable Fundamental Cacheability for Distributed Deep Learning Training**
Invited talk: The GenAI and ML Systems Efficiency Workshop, Adobe Research, virtual (10/2023)

2023 **Stateful Computing in a Serverless Way**
Invited talk: The University of Edinburgh, Scotland, virtual (04/2023)

2022 **Computing in a Serverless Way for Fun and Profit**
Invited talk: Virginia Tech Northern Virginia Center, Falls Church, VA (10/2022)

2022 **Scaling Data Analytics on Serverless Clouds**
Invited talk: McDaniel College, MD (03/2022)

2018 **Analyzing Alibaba's Co-located Datacenter Workloads**
Conference talk: IEEE BigData 2018, Seattle, WA (12/2018)

2018 **The hardware, they are a-changin**
Breakout summary talk: Workshop on Data Storage Research 2025, San Jose, CA (05/2018)

2018 **Breaking the Monolith: Rethinking Storage System Design**
Invited talk: Virginia Tech Northern Virginia Center, Falls Church, VA (03/2018)

2018 **Erasing Belady's Limitations: In Search of Flash Cache Offline Optimality**
Invited talk: HPDC'18 TPC Workshop, Berkeley, CA (03/2018)

2017 **Breaking the Monolith: Rethinking Storage System Design**
George Mason University, Fairfax, VA (11/2017)
George Mason University, Fairfax, VA (04/2017)

2016 **Erasing Belady's Limitations: In Search of Flash Cache Offline Optimality**
Conference talk: USENIX ATC'16, Denver, CO (06/2016)
Internship talk: The CTO Office of EMC CTD, Princeton, NJ (06/2016)

2015 **Pricing Games for Hybrid Object Stores in the Cloud: Provider vs. Tenant**
Conference talk: USENIX HotCloud'15, Santa Clara, CA (06/2015)
The CTO Office of EMC CTD, Princeton, NJ (05/2015)

2015 **CAST: Tiering Storage for Data Analytics in the Cloud**
Conference talk: ACM HPDC'15, Portland, OR (06/2015)

2015 **An In-Memory Object Caching Framework with Adaptive Load Balancing**
Conference talk: ACM EuroSys'15, Bordeaux, France (04/2015)

2014 **An In-Memory Object Caching Framework with Adaptive Load Balancing**
Internship talk: IBM Almaden Research Center, San Jose, CA (08/2014)

2013 **High Performance, Flexible Memory Caching**
Internship talk: IBM Almaden Research Center, San Jose, CA (08/2013)

Teaching

At University of Virginia

Spring 2024 **CS/DS5110 Big Data Systems**
Enrollment: 100

Spring 2023 **DS5110 Big Data Systems**
Enrollment: 64

At George Mason University

Spring 2022 **CS571 Operating Systems**
Enrollment: 23, —Overall instructor rating and course rating cancelled starting Spring 2022—

Fall 2021 **CS475 Concurrent & Distributed Systems**
Enrollment: 58, Instructor rating: 4.36/5, course rating: 4.16/5

Spring 2021 **CS571 Operating Systems**
Enrollment: 18, Instructor rating: 4.93/5, course rating: 4.64/5

Fall 2020 **Teaching leave**

Spring 2020 **CS675 Distributed Systems**
Enrollment: 9 (formal teaching evaluation cancelled due to COVID-19)

Spring 2020 **CS571 Operating Systems**
Enrollment: 34 (formal teaching evaluation cancelled due to COVID-19)

Fall 2019 **CS471 Operating Systems**
Enrollment: 68, Instructor rating: 4.33/5, Course rating: 3.98/5

Spring 2019 **CS471 Operating Systems**
Enrollment: 66, Instructor rating: 4.63/5, Course rating: 4.06/5

Fall 2018 **CS795 Cloud Computing**
Enrollment: 8, Instructor rating: 4.88/5, Course rating: 4.88/5

Fall 2017 **CS471 Operating Systems**
Enrollment: 59, Instructor rating: 2.94/5, Course rating: 2.81/5

Student Advising

PhD Dissertation Advisor

1. Zheng Chai, PhD, CS@UVA, *8 papers published, 1 paper under review*, started 2018, expected to graduate Spring 2024
   Topic: Distributed machine learning systems
   Internships:
   - HPE, Summer 2021.
2. Yuqi Fu, PhD, CS@UVA, *1 paper published* started 2020
   Topic: Serverless resource scheduling
   Internships:
   - ByteDance, Summer 2022.
3. Benjamin Carver, PhD, CS@GMU, *2 papers published*, started 2021
   Topic: Stateful serverless computing
   Internships:
   - Microsoft Research, Summer 2022.
4. Zhaoyuan (Alex) Su, PhD, CS@UVA, *1 paper published*, started 2021
   Topic: Algorithmic and systems support for large-scale federated learning
   Internships:
   - Argonne National Laboratory, Summer 2022.
5. Rui Yang, PhD, CS@UVA, started 2021
   Topic: Learned data storage systems

PhD Graduates

1. Jingyuan Zhang, PhD, CS@GMU, 2018–2023
   Dissertation: Towards Elastic and Cost-effective Stateful Serverless Systems
   First employment: Cloud native infrastructure team @ ByteDance

Master Research

1. Benjamin Carver, Accelerated BS/MS Program@GMU, *2 papers published*
   Topic: Designing a Serverless Data Analytics Framework
2. Rafael Madrid MS, CS,
   Topic: Designing NVM Storage for Serverless Workloads
3. Anne Martine Augustin (MS, SWE, Spring'19–Summer'19)

Undergraduate Research

Shengming Gao, CS@UVA

Michael Somarriba, CS@GMU

Daniel Meneses, CS@GMU

Yuanqi Du, CS@GMU

Benjamin Carver, CS@GMU

Isaiah King, CS@GMU

Dawen Yang, CS@GMU

Mark Boehen, ECE@GMU

Hannan Fayyaz, CS, York University, Canada

Zeshan Fayyaz, CS, Ryerson University, Canada

PhD Dissertation Committee Member

Redwan Ibne Seraj Khan, PhD, CS@VT
Samuel S. Ogden, PhD, CS@WPI
Hengrun Zhang, PhD, CS@GMU
Li Liu, PhD, CS@GMU

Professional Services

University, College, and Department Service

2021–2022 Faculty search committee, Computer Science, GMU
2017–2019 Ph.D. admissions committee, Computer Science, GMU

Conference Organizer and Community Services

2024 **HotStorage**, General co-chair, ACM Workshop on Hot Topics in Storage and File Systems
2024-2023 **HPDC**, Workshop co-chair, ACM International Symposium on High-Performance Parallel and Distributed Computing
2023-2022 **HotStorage**, Publication chair, ACM Workshop on Hot Topics in Storage and File Systems
2021-present **IEEE STCOS**, Co-chair, IEEE Special Technical Community on Operating Systems
2021 **ICDCS**, Local arrangement chair, IEEE International Conference on Distributed Computing Systems
2019 **SEC**, Local arrangement chair, ACM/IEEE Symposium on Edge Computing

Editorial Boards

2023-present **Associate Editor**, Frontiers in High Performance Computing: Cloud Computing

Technical Program Committee

2024 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing
2024 **IPDPS**, IEEE International Parallel and Distributed Processing Symposium
2023 **SoCC**, ACM Symposium on Cloud Computing
2023 **HotStorage**, ACM Workshop on Hot Topics in Storage and File Systems
2023 **IEEE Cloud**, IEEE International Conference on Cloud Computing
2023 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing
2023 **IPDPS**, IEEE International Parallel and Distributed Processing Symposium
2022 **NAS** (storage track), IEEE International Conference on Networking, Architecture, and Storage
2022 **KDD** (ERC), ACM SIGKDD International Conference on Data Mining
2022 **HiPS**, Workshop on High Performance Serverless Computing@HPDC 2022
2022 **SEC**, ACM/IEEE Symposium on Edge Computing
2022 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing
2021 **REX-IO**, Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads
2021 **ICDCS**, 41^{st} IEEE International Conference on Distributed Computing Systems
2021 **SEC**, ACM/IEEE Symposium on Edge Computing
2021 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing
2020 **PDSW-DISCS**, 5^{th} International Parallel Data Systems Workshop
2020 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing
2020 **ICDCS**, 40^{th} IEEE International Conference on Distributed Computing Systems
2020 **SC**, International Conference for High Performance Computing, Networking, Storage, and Analysis
2020 **MSST**, 36^{th} International Conference on Massive Storage Systems and Technology
2020 **CCGrid**, IEEE/ACM International Symposium in Cluster, Cloud, and Grid Computing
2019 **PDSW-DISCS**, 4^{th} International Parallel Data Systems Workshop

2019 **MASCOTS**, 27^{th} IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems

2019 **IPDPS** (ERC), IEEE International Parallel and Distributed Processing Symposium

2019 **CCGrid** (ERC), IEEE/ACM International Symposium in Cluster, Cloud, and Grid Computing

2019 **BlockDM**, First IEEE International Workshop on Blockchain and Data Management

2019 **MSST**, 35^{th} International Conference on Massive Storage Systems and Technology

2019 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing

2018 **HPDC**, ACM International Symposium on High-Performance Parallel and Distributed Computing

2018 **ICS** (ERC), ACM International Conference on Supercomputing

2018 **IPDPS** (ERC), IEEE International Parallel and Distributed Processing Symposium

2018 **ICCCN**, International Conference on Mobile Systems and Pervasive Computing

2018 **MobiSPC**, International Conference on Computer Communications and Networks

2017 **BDCAT**, IEEE/ACM International Conference on Big Data Computing, Applications and Technologies

Proposal Review Panels

2023 **DOE**, Office of Science, Advanced Scientific Computing Research (ASCR) Program

2021 **NSF**, Computer Systems Research (CSR) under the division of Computer and Network Systems (CNS)

2020 **NSF**, Computer Systems Research (CSR) under the division of Computer and Network Systems (CNS)

2019 **NSF**, Computer Systems Research (CSR) under the division of Computer and Network Systems (CNS)

2019 **NSF**, Software and Hardware Foundations (SHF) under the division of Computing and Communication Foudations (CCF)

Shadow Technical Program Committees

2018 **EuroSys**, ACM European Conference on Computer Systems

2017 **EuroSys**, ACM European Conference on Computer Systems

2016 **EuroSys**, ACM European Conference on Computer Systems

Journal Reviews

2023-2017 **TOS**, ACM Transactions on Storage

2022-2018 **TCC**, IEEE Transactions on Cloud Computing

2022 **TNNLS**, IEEE Transactions on Neural Networks and Learning Systems

2022-2017 **JPDC**, Journal of Parallel and Distributed Computing

2022 **TMC**, IEEE Transactions on Mobile Computing

2019-2017 **TC**, IEEE Transactions on Computers

2020-2015 **TPDS**, IEEE Transactions on Parallel and Distributed Systems

2017 **TAAS**, ACM Transactions on Autonomous and Adaptive Systems