Yue Cheng — Examples of Collaboration Statement

I request consideration for promotion to Associate Professor with tenure in the School of Data Science at the University of Virginia (UVA). This statement describes examples of my collaborations with faculty, staff, students, and industry in scholarship and research activities during my last five years at George Mason University (GMU) and my agenda for continued collaborative work at UVA for years to come.

I have been collaborating with faculty, staff, industry companies, and national laboratories for over a decade in research and service activities. In fact, after I started my tenure-track faculty career, most of my research work is collaborative. I will describe these collaboration activities next and emphasize if this collaboration is led by me and/or my research group shares a dominant credit of the collaborative work.

1 Collaboration with Faculty

I have had successful collaborations with faculty from George Mason University (GMU), Emory University, Auburn University, University of Minnesota (UMN), University of Nevada, and Virginia Tech (VT).

My collaboration with GMU faculty Prof. Songqing Chen has produced three quality research papers that appeared (and to appear) at Computer Science conferences such as Virtual Execution Environments (VEE) [VEE'19], Symposium on Cloud Computing (SoCC) [SoCC'21], and the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC or SuperComputing) [SC'22]. Prof. Songqing Chen was the lead for the first two papers (VEE and SoCC), while I am the project lead for the third paper (SC). In the SC work where I am the team lead, I propose a new operating system (OS) scheduling algorithm that can improve the (machine learning and data analytics) application performance by two to three orders of magnitude, compared to the existing OS scheduler Completely Fair Scheduler (CFS). This work has also been nominated as one of the five **Best Student Paper Award Finalists** (one in more than 500 papers submitted to SC'22) in the prestigious SC 2022 conference.

I have been collaborating with Prof. Liang Zhao from Emory University for over three years. Our collaboration has been fruitful: our collaborative project has won an NSF OAC Small grant ($500k) [NSF OAC], an Amazon Research Award ($75k Amazon Web Service Cloud credit) [Amazon Research Award], and a recent Meta (previously Facebook) Research Award ($50k industry gift) [Meta Research Award]. Prof. Liang Zhao was the lead for the NSF OAC Core project and Amazon Research Award project, while I am the lead for the Meta Award project. In the Meta Award project in collaboration with Meta, I propose a novel, compute-and-storage-disaggregated training architecture for training emerging, large graph neural networks (GNNs). In addition to the various grants we have secured, this collaboration has also produced a line of quality research papers in top venues including the IEEE International Conference on Data Mining (ICDM) [ICDM'20] and the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC or SuperComputing) [SC'21].

As another example, I have been collaborating with Prof. Peter Liu from Auburn University on an interdisciplinary project that connects advanced additive manufacturing with the emerging federated learning techniques. We have secured a two-year, $500k NSF FMSG grant (Future Manufacturing Seed Grant) [NSF FMSG] to support our interdisciplinary research effort. It is exciting to see how advanced data analytics techniques (federated learning) and data-driven research methodologies can be applied and adapted to domain engineering fields such as manufacturing and 3D-printing.

I am also actively collaborating with UVA faculty in Data Science (Prof. Sheng Li) and Computer Science (Prof. Haiying Shen, Prof. Tianhao Wang). Our collaboration at UVA will hopefully see its fruit

in publications and fund raising soon. I am also planning to further extend my collaboration with other faculty both with the School of Data Science and outside to further the research impact.

2 Collaboration with Staff

I have been actively collaborating university staff on research and service activities. For example, I have collaborated with Dr. Jayshree Sarma, the director of GMU's Office of Research Computing (ORC), on designing and constructing Hopper [GMU Hopper Project], a high-performance research instrument that supports GMU's high-impact multidisciplinary research and education. We have secured a competitive NSF project *MRI-2018631: MRI: Acquisition of an Adaptive Computing Infrastructure to Support Compute- and Data-Intensive Multidisciplinary Research* ($750k) [NSF CRI] to build the HPC cluster on GMU's campus.

3 Collaboration with Industry and National Laboratories

I have maintained close collaboration relationships with leading industry companies including Adobe, Alibaba, Microsoft, Meta (Facebook), IBM, Amazon, and Western Digital, and national laboratories including Argonne National Laboratories (ANL). These industrial collaborations are in various forms including sending my Ph.D. students for summer research internships and collaborating with industrial researchers on research problems in production. Next I will briefly describe the collaborations with Adobe, Alibaba, Microsoft, and ANL.

I have been collaborating with Adobe Research on designing the next-generation, serverless GPU notebook cloud service for Adobe Sensei [Adobe Sensei], which is Adobe's ML (machine learning) and AI (artificial intelligence) platform. In collaboration with Dr. Kanak Mahadik and Dr. Haoliang Wang from Adobe Research, I am designing ElasticAI, a cloud-native, co-designed, serverless GPU and storage scheduler will be developed to evaluate our approach. This research is partially supported by a contiguous Adobe research gift with a total amount of $50k to date. We are in the final wrap-up phase of this project and we hope to submit a research paper to a top ML systems conference in the near future. We have already demo'ed a developed prototype within Adobe and the results suggest that, once adopted by Adobe, our technique can save up to around 3 million dollars of cloud resource cost for Adobe Sensei.

I also have successful experience collaborating with Alibaba Cloud. This collaboration started with sending my Ph.D. student to Alibaba Cloud for a summer research internship. In this project, we helped design a novel distributed data structure called a *Function Tree*, which effectively decentralizes the burdensome provisioning process of large container runtimes to a large number of virtual machines (serving as nodes of the FaaS tree). The resulted system, which we call FaaSNet [ATC'21], has been adopted and integrated into the service infrastructure of Alibaba Cloud Function Compute [Alibaba Cloud Article], and has been constantly serving millions of user requests on a daily basis.

My Ph.D. student did a summer research internship at Microsoft Research in Summer 2022, working with Dr. Rodrigo Fonseca (the Principal Researcher at Microsoft who lead the Azure Systems Research (AzSR) group) on building Microsoft's next-generation stateful serverless computing platforms. This research is motivated by my research findings on the Wukong project [SoCC'20], where we built a highly-efficient serverless data analytics engine on Amazon Web Services. We are currently implementing the Wukong idea into Azure's in-house stateful serverless computing platforms and are preparing a manuscript that will be submitted to a top Computer Science conference by the end of 2022.

We have been collaborating with research scientists Dr. Sheng Di and Dr. Frack Cappello from the Argonne National Laboratories on designing a novel lossy compression algorithm for scientific datasets. In this collaboration, we focus on understanding the impact of various error-controlled lossy compressors on multiple derivative-related metrics commonly concerned by users. We perform solid experi-

ments that involve 5 state-of-the-art lossy compressors and 4 real-world application datasets (primarily for scientific data visualization).

4 Concluding Remarks

This statement summarizes examples of my collaboration with faculty, staff, companies, and national labs. I am grateful for these fantastic collaboration opportunities and I hope to further extend collaborations with leading faculty and organizations by generating highly-impactful scholarly work for the years to come.

Selected Publications, Projects, Grants, and Medium Featuring

[SoCC'20] Benjamin Carver, Jingyuan Zhang, Ao Wang, Ali Anwar, Panruo Wu, **Yue Cheng**. WUKONG: *A Scalable and Locality-Enhanced Framework for Serverless Parallel Computing*, In ACM SoCC '20.

[ATC'21] Ao Wang, Shuai Chang, Huangshi Tian, Hongqi Wang, Haoran Yang, Huiba Li, Rui Du, **Yue Cheng**. FAASNET: *Scalable and Fast Provisioning of Custom Serverless Container Runtimes at Alibaba Cloud Function Compute*, In USENIX ATC '21.

[SoCC'21] Li Liu, Haoliang Wang, An Wang, Mengbai Xiao, **Yue Cheng**, Songqing Chen. *Mind the Gap: Broken Promises of CPU Reservations in Containerized Multi-tenant Clouds*, In ACM SoCC '21.

[SC'22] Yuqi Fu, Li Liu, Haoliang Wang, **Yue Cheng**, Songqing Chen. *SFS: Smart OS Scheduling for Serverless Functions*, In SC '22.

[VEE'19] Li Liu, Haoliang Wang, An Wang, Mengbai Xiao, **Yue Cheng**, Songqing Chen. *vCPU as a Container: Towards Accurate CPU Allocation for VMs*, in ACM VEE '19.

[SC'21] Zheng Chai, Yujing Chen, Ali Anwar, Liang Zhao, **Yue Cheng**, Huzefa Rangwala. *FedAT: A High-Performance and Communication-Efficient Federated Learning System with Asynchronous Tiers*, In SC '21.

[NSF FMSG] *CMMI-2134689: FMSG: Cyber: Federated Deep Learning for Future Ubiquitous Distributed Additive Manufacturing*, https://www.nsf.gov/awardsearch/showAward?AWD_ID=2134689, 2021.

[NSF CRI] *MRI: Acquisition of an Adaptive Computing Infrastructure to Support Compute- and Data- Intensive Multidisciplinary Research*, https://www.nsf.gov/awardsearch/showAward?AWD_ID=2018631&HistoricalAwards=false.

[ICDM'20] Junxiang Wang, Zheng Chai, **Yue Cheng**, Liang Zhao. *Toward Model Parallelism for Deep Neural Network based on Gradient-free ADMM Framework*, In IEEE ICDM '20.

[NSF OAC] *OAC-2007976: OAC Core: SMALL: DeepJIMU: Model-Parallelism Infrastructure for Large-scale Deep Learning by Gradient-Free Optimization*, https://www.nsf.gov/awardsearch/showAward?AWD_ID=2007976, 2020.

[Amazon Research Award] *Amazon Research Award: Distributed Large-scale Graph Deep Learning by Gradient-free Optimization*, 2020.

[Meta Research Award] *Serverless and Scalable GNN Training with Disaggregated Compute and Storage*, https://research.facebook.com/research-awards/2022-request-for-research-proposals-for-ai-system-hardware-software-codesign/, 2022.

[NSF SPX] *CCF-1919075: SPX: Collaborative Research: Cross-stack Memory Optimizations for Boosting I/O Performance of Deep Learning HPC Applications*, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1919075, 2019.

[GMU Hopper Project] https://content.sitemasonry.gmu.edu/news/2020-09/need-speed-mason-makes-huge-strides-

[Adobe Sensei] https://www.adobe.com/sensei.html.

[Alibaba Cloud Article] https://www.alibabacloud.com/blog/597937.

[Wukong Project] https://ds2-lab.github.io/Wukong/.

[FaaSNet Project] https://github.com/ds2-lab/FaaSNet.

[SFS Project] https://github.com/ds2-lab/SFS.