Serverless computing or Function-as-a-Service (FaaS) is gaining traction because a key goal of serverless computing is simplicity: there are almost no configuration knobs for users to configure. Serverless computing platforms can scale up to thousands of threads in seconds, which run virtually anything a cloud user supplies as a serverless function. Once the functions finish execution, they scale down to zero. A fundamental question to ask is, how can we support large, stateful applications (e.g., complex analytics jobs and storage services) using FaaS? In fact, running these stateful jobs on today’s FaaS platforms is inefficient and programming these applications in a serverless way is even more difficult. In this talk, I would like to make a strong case for scaling large, data-intensive applications on serverless clouds. I will introduce two works that our group has been working on in the past several years: 1. scaling Python data analytics on FaaS platforms, and 2. supporting cost-effective data storage services on FaaS platforms. In the first work, I will show you how Wukong, a serverless parallel computing framework, can help data scientists program massive CPU and memory resources at any scale by using FaaS; in the second work, I will show you how InfiniCache, a first-of-its-kind serverless in-memory caching system, helps reduce the cloud storage cost by two orders of magnitude with FaaS. Finally, I will briefly talk about our recent efforts in designing an end-to-end serverless computing stack spanning from the application layer all the way to the OS layer.

Biography

Dr. Yue Cheng is an assistant professor of Data Science and Computer Science at the University of Virginia (UVA). He received his Ph.D. in Computer Science from Virginia Tech in 2017. He was an assistant professor of Computer Science at George Mason University (GMU) from 2017 to 2022. His research interests include distributed systems, cloud and
serverless computing, and high-performance computing. His current research focuses on the systems support for cloud- and HPC-scale data-intensive computing. His research results have been published at major conferences such as SC, HPDC, EuroSys, ATC, SoCC, and ICDM. He has developed a set of new techniques that make stateful serverless computing efficient, scalable, and easy to program. His work on serverless data analytics has been used by tech giants including Alibaba, Microsoft, and Adobe. Dr. Yue Cheng is the recipient of an NSF CAREER Award, a Meta Research Award, an Amazon Research Award, and the 2022 IEEE CS TCHPC Early Career Researchers Award for Excellence in High Performance Computing.