Statement of Purpose

The escalating need for computing resources has been supported by the rapid development of the cloud computing paradigm. Different tiers of the system stack have predominantly evolved independently. However, poor coordination across these levels—from architectural design to operating systems (OS) and cloud-native applications—has resulted in numerous inefficiencies in resource utilization and scaling. My research interests focus on OS and cloud computing. Specifically, I aim to enhance core OS subsystems of CPU scheduling, memory, and energy management tailored toward creating strong synergies for cloud applications. Through a Ph.D. program in Computer Science at the University of Illinois, Urbana-Champaign, I wish to make fundamental contributions to my overarching research goal of harnessing the full potential of the cloud.

Over fifteen years after the introduction of the cloud, utilizing resources efficiently remains an ongoing challenge. The surge in migrating traditional workloads and the emergence of newer ones, such as machine learning applications, have made this problem over the infrastructure more pertinent than ever. Currently, as an MS student at the University of Illinois, Urbana-Champaign (UIUC), my focus lies on addressing a fundamental disconnect between the operating system's CPU entitlement interface and the resource requests of cloud applications. Applications historically seek resources in terms of full or partial cores, while the OS interfaces rely on CPU allocation as a function of bandwidth. The state-of-the-art auto scalers fail to consider application behavior within this CPU allocation paradigm, resulting in sluggish, inaccurate, and disruptive entitlement recommendations. The problem is worsened further for serverless applications which display shorter lifespans, minimal heterogeneity, and require higher reactivity. Therefore as part of my MS thesis, collaborating with Professors Saugata Ghose and Tianyin Xu, I am currently working on designing a light-weight tracing framework to analyze a cgroup's CPU bandwidth utilization behavior within the OS and recommend the correct CPU entitlement with high degrees of reactivity. Initial results suggest gains in both performance and CPU efficiency for various cloud workloads.

In a separate research project at UIUC with Prof. Ghose, I'm delving into relationships that extend beyond spatial and temporal connections among data entities. Traditionally, applications commonly exhibited locality, yet many vital modern applications such as databases and graph processing lack high degrees of this behavior. Consequently, their performance suffers, accompanied by substantial energy wastage. In pursuit to address this challenge, we've identified relationships of clustered objects that repeatedly interact with each other but are not captured by traditional locality paradigms. I am also currently working in collaboration with an undergraduate student as part of the Illinois Undergraduate Scholar Research (ISUR) program. Our goal is twofold: firstly, to effectively quantify the degree of this relationship within contemporary workloads; and secondly, to develop memory management policies. These policies will encompass allocation, eviction, and prefetching strategies across architecture and operating systems, aiming to mitigate the data movement bottleneck. I hope to submit first-author papers for the above two research endeavors by my MS graduation.

Prior to graduate school at UIUC, I worked as a kernel engineer at IBM - Linux Technology Center for three years. I began as a six month intern and contributed to enhancing the gem5 architecture simulator to support the IBM POWER architecture. This effort culminated in the successful booting of a multi-threaded Linux kernel onto the simulator and was presented at the OpenPOWER Summit 2019. Transitioning into a full-time kernel engineer, my primary responsibilities revolved around contributing in the area of scheduling and energy management for Linux. Notably, I spearheaded a project in designing a probabilistic approach to choose CPU-idle states that both helped save energy and boosted performance. Presenting this feature at the OS-Directed Power-Management (OSPM) Summit 2020, Italy, and its coverage in the Linux Weekly News marked significant milestones for me. My contributions with IBM also included the area overlapping CPUs and container primitives. I proposed CPU namespace - a mechanism to virtualize topology information for containers, resulting in significant efficiency gains and performance isolation for cloud workloads. My proposal was published by Phoronix and was accepted at the Linux Conference Australia (LCA) 2022. Additionally, to recognize my efforts, I was awarded the General Manager Awards in the Exemplary Rookie category - awarded to one employee across divisions.

My research interest in systems predates my time at UIUC and IBM, stemming from an undergraduate summer internship at Carnegie Mellon University (CMU) with Professor Saugata Ghose. During this period, I was involved in profiling memory patterns from the operating system's perspective. Additionally, I designed an OS memory access simulator to analyze these patterns and evaluate various memory prefetching algorithms. In addition to my research pursuits, I also enjoy teaching and mentoring. I have served as a teaching assistant for CS 233 computer architecture at UIUC for a year. Additionally, I've taken on mentoring roles, guiding students both at IBM and UIUC (ISUR Program). These experiences have further solidified my commitment to nurturing emerging talent within the field.

My background across architecture, operating systems, and cloud computing from CMU, IBM, and UIUC has provided me with a distinctive perspective on the practical hurdles in developing efficient cloud systems. At UIUC, I am exceedingly fascinated with projects at Prof. Saugata Ghose’s ARCANA Research Group. The innovative strides taken toward developing data-centric computing and the vision for enabling novel architectures such as Processing-in-Memory through hardware software co-design have significantly influenced me. Simultaneously. Prof. Tianyin Xu’s work on building reliable systems for the cloud such as correctness verification of cloud operators, introducing Rust to mitigate the limitations of the in-kernel eBPF verifier, and a mechanism to concurrently execute multiple invocations of serverless functions align with my research aspirations. Likewise, Prof. Deepak Vasisht’s endeavors in improving networking techniques for satellites in space hold significant intrigue to me. My research interests in improving scheduling, memory management, and energy efficiency at the intersection of operating systems and cloud computing resonate with the direction that UIUC is pioneering as well. I believe that the Ph.D. program will greatly bolster my journey in shaping high-performing and efficient cloud computing paradigms of the future.