As a member of generation Z, I grew up during China's Internet boom. The Internet is fundamental to our lives. As a computer scientist, I aspire to push the Internet to evolve in the next few decades.

During my undergraduate studies at ShanghaiTech University and the University of Illinois at Urbana-Champaign, I engaged in innovative research. I strive to build an operating system and network layers that require no human intervention and will run any task without errors for decades. Designing and building an entire system from the hardware up and to the networking layers is challenging and requires years of studies and research. Therefore, my mission as a graduate student is to engage with cutting-edge research, to find and solve new problems, and move step by step to implement this ideal system.

Computing infrastructure is a contemporary industry focus. During a visit to a local big data center in Ya'an, China, the operators there indicated that they hope to revitalize the local economy by providing hosting services for tech companies. When I asked them why they couldn't make more profit by selling more diverse services, such as serverless service, they told me that there was a lack of expertise. Therefore, through my future work, I want to make it easier for more people to enjoy the benefits of computers and the Internet. I identify three challenges that stand before this goal. Over the last few years, my research projects involved the three challenges. First, merging IO-intensive applications require a unified and flexible way to abstract the system's hardware and software resources. Second, human-introduced errors unavoidable in contemporary systems and cloud services need to become more intelligent and compact. Third, as Moore's Law reaches ever closer to obsolescence, we need to embrace heterogeneous architecture to maximize efficiency gains with lower costs.

I started my first research project in 2021 with Professor Zhice Yang at ShanghaiTech University. During the COVID-19 pandemic, instructors had difficulties obtaining specialized equipment to annotate slides during online classes. To address this problem, we used commodity devices like Bluetooth headsets and cell phones to implement gesture tracking through acoustic localization. With our innovation, the instructor can hold their Bluetooth headset in their hands and draw annotations on the wall. We can track the instructor's gesture by calculating the relative position between the cell phone and the Bluetooth headsets. We calibrated the clock on both devices via Bluetooth to improve the accuracy of the acoustic localization. We achieved a clock synchronization accuracy of 40ms. From this project, I had my first taste in research and I find myself amazed by the incredible work unassuming devices can do.

The second research started in 2022. While people are using mobile applications to scan QR codes, sensitive information like home environment can be captured unintentionally. We want to remove sensitive information from the information stream, such as blurring the background except the QR code, to achieve fine-grained privacy control. We modified the camera API in Android framework to implement region of interest detection with OpenCV. This experience gave me exposure to the underlying system code and building systems.

In 2022, as a team member under Professor Shu Yin from ShanghaiTech, I participated in the ISC 2022 Student Cluster Competition and took fourth place. For our competition, we were tasked to optimize ICON, a weather forecasting program. I profiled and analyzed the environment and improved the program's performance by decreasing MPI wait and increasing CPU affinity. The optimized ICON was 30% faster than the baseline.

Stepping beyond student competition, my research focused on resolving real-world HPC challenges. Working with Professors Shu Yin and Zhiling Lan, our team conducted an in-depth study on HPC-AI workloads. We found that the existing high performance computing cluster' scheduler could not handle the growing demand for HPC-AI applications because the best performing machines are always overloaded with AI tasks. We took a profile of the GPU kernel usage and an implementation of an AI scheduler. With our AI-Scheduler, the AI programs with low GPU demand are scheduled to low-end but highly available machines, whereas some powerful GPUs will be reserved for incoming high-load tasks. The response time of a high-performance cluster could then be reduced. We achieved a 10% decrease in response time.

In Fall 2022, I came to the University of Illinois at Urbana-Champaign as an exchange student to work with Professors Tianyin Xu and Jovan Stojkovic. Existing serverless implementations perform poorly with high concurrent requests and take up too many resources. We wanted to save resources and improve response times by using our optimized MXContainer. My work is to migrate MXContainer, a highly optimized serverless platform to KNative. The MXContainer employs novel techniques, such as IO-sharing, CPU-sharing, and memory-sharing to serverless and improves the performance while reducing the application's cost. The KNaitve is a widely used serverless platform and we aimed to optimize it. We decreased the tail latency of serverless applications by 50% and eased the system resources utilization. We aim to push our work upstream and have help from IBM/RedHat serverless teams. This experience exposed me to frontier serverless platforms and had me work with more intelligent people. I am excited about the research in operating systems and networks.

What's more, I committed myself to the open source software movement and other interesting things. I am a member of the Anthon Open Source Community and NekomimiRouter. Through our work in the Anthon Open Source Community, I can proudly use our own Linux distributions daily. I improved its usability by software packaging. NekomimiRouter is a community made of people who want to make the Internet better. In NekomimiRouter, we are trying to build our own network infrastructure. I can access the Internet through a self-maintained ISP.

Thank you for your time and consideration.