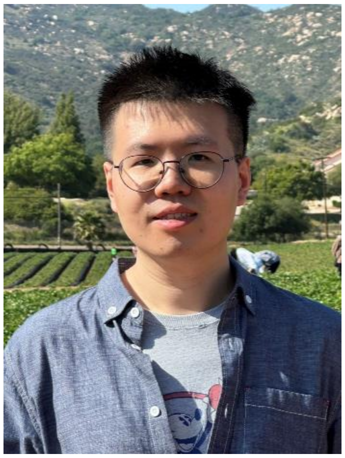
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

# DISSERTATION DEFENSE



## Yiwen Zhang

### Quality of Service for Performance-Critical Cloud Applications

Tuesday, July 9, 2024
1:00pm – 3:00pm
4941 Beyster
Hybrid – Zoom

**ABSTRACT:** Cloud infrastructure continues to scale due to rapid evolvement of both hardware and software technologies in recent years. As a result, modern cloud applications are becoming increasingly fast and efficient. However, in a shared environment, multiple applications with different performance requirements must coexist and share cloud resources. Such sharing maximizes cloud utilization, but also leads to resource contention, causing unpredictable performance. Therefore, it becomes extremely important to ensure performance-critical applications receive the appropriate level of priority and service quality.

This dissertation aims to build system support for better quality of service (QoS) for performance-critical applications in the cloud. Specifically, I aim to provide guaranteed performance specified by service level objectives (SLOs) for multiple coexisting applications while maximizing resource utilization. Over the past few years, I have built several systems to provide QoS among various places in the cloud -- including host network interface card, datacenter networks, edge infrastructure, and computer memory systems. In this talk, I will introduce four research projects in my thesis and share ideas on (1) providing performance isolation in kernel bypass networks, (2) designing admission control for performance-critical datacenter RPCs, (3) performing automatic query planning for live ML analytics, and (4) building QoS-aware multi-tiered memory systems.

**CHAIR:** Prof. Mosharaf Chowdhury