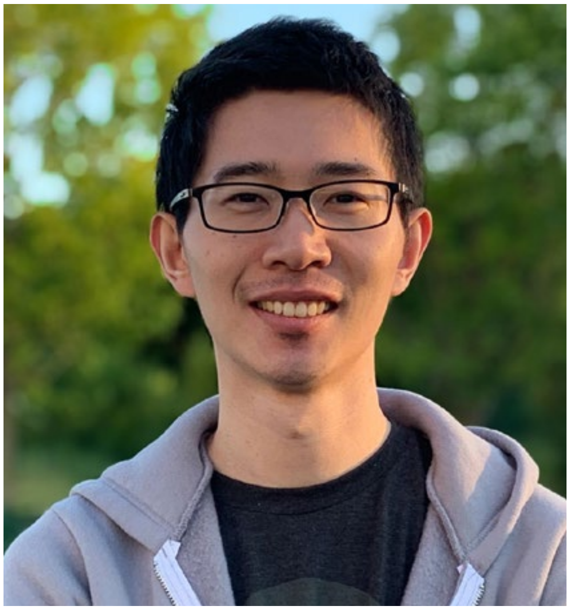
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



# Yibo Pi

## Evaluating and Improving Internet Load Balancing with Large-Scale Latency Measurements

Thursday, March 4, 2021
12:00 PM
Virtual (Passcode: 545845)

**ABSTRACT:** Load balancing is used in the Internet to distribute load across resources at different levels, from global load balancing that distributes client requests to respective nearby servers at the Internet level to path-level load balancing that balances traffic across load-balanced paths. These load balancing algorithms generally work under certain assumptions on performance similarity. Specifically, global load balancing divides the Internet address space to client aggregations and assumes that clients in the same aggregation have similar performance to the same server, and load-balanced paths are generally selected for load balancing as if they have similar performance. However, as performance similarity is typically achieved with similarity in path property, e.g., topology and hop count, which does not necessarily lead to similar performance, this could cause significantly different performance between clients in the same aggregation and between load-balanced paths.

This dissertation evaluates and improves global and path-level load balancing in terms of performance similarity. We achieve this with large-scale latency measurements, which not only allow us to systematically identify and evaluate the performance issues of Internet load balancing at scale, but also enable us to develop data-driven approaches to improve its performance. Specifically, this dissertation consists of three parts. First, we study the issues of existing client aggregations for global load balancing and then design AP-atoms, a data-driven client aggregation learned from passive large-scale latency measurements. Second, we show that the latency imbalance between load-balanced paths, previously deemed insignificant, is now both significant and prevalent. We design Flipr, a network prober that actively collects large-scale latency measurements to characterize the latency imbalance issue. Lastly, we design another network prober, Congi, that can detect congestion at scale and use Congi to study the congestion imbalance problem at scale. We further demonstrate that latency and congestion imbalance could greatly affect the performance of various applications.

**CHAIR:** Prof. Sugih Jamin