"Optimizing Webservers With io_uring"

By Mihika Nigam

Abstract:

Modern webservers face unprecedented demands for high throughput and low latency, yet even the state-of-the-art optimizations often fail under heavy workloads on our existing hardware. Despite advancement in hardware, communication between applications and the kernel remains a critical bottleneck.

Industry surveys reveal over 50% of production servers still rely on traditional epoll-based architectures [5]. This research aims to investigate and characterize Linux's new io_uring subsystem that can overcome these challenges. With controlled load testing of all the existing architectures like epoll, multi-process, and multi-threaded architectures (including other commercial servers), we demonstrate how io_uring can help achieve better benchmarks with and without cached data. Our methodology employs controlled load testing while maintaining constant network bandwidth as the primary bottleneck, with detailed measurement of key system metrics including syscalls, memory, disk I/O performance and CPU utilization patterns.

This work provides empirical guidance for migrating performance-critical services to io_uring while outlining remaining challenges for future work.

Keywords: WebServers, IO_URING, Jmeter, O_DIRECT

Date: Monday, June 9th, 2025

Time: 3:00 PM – 5:00 PM

Location: 14-238b

Committee: Dr. Bellardo, Dr. Beard, Dr. Schmitt