Lai Wei, Ph.D.

Phone: (832) 312-7308
Email: laiweicq@waymo.com
Website: https://laiwei.weebly.com/

Summary

Compute workload specialist in Waymo's Compute team. Responsible for analyzing AV (autonomous vehicle) workloads and improving hardware/software stack to achieve cross-platform top performance. Worked at Compute Performance team of Pony.ai, and performance optimization in HPC (High Performance Computing).

- Experience with performance analysis and optimization of software.
  - Experience with performance tools, e.g. perf, gperftools, VTune, eBPF, HPCToolkit, etc.
  - Experience with parallel programming models, e.g. MPI, OpenMP, CUDA, Cilk, etc.
  - Experience with performance optimization in sophisticated software systems.
- Knowledge of computer architecture, assembly, compiler, operating system, etc.
- Languages: C++, Python, Java.

Past Experience

Aug 2022 - Present • Compute Workload Specialist • **Waymo** • Mountain View, California

- Identify and perform deep dives into important AV workloads.
- Understand and help define the next-generation hardware/software architecture.

Oct 2019 - Aug 2022 • Senior Software Engineer • **Pony.ai** • Fremont, California

- Monitored, analyzed, and optimized performance of AV software.
  - Collected and monitored key performance metrics of AV software system.
  - Developed and maintained performance analysis tools for AV workloads.
  - Optimized source code and libraries within AV software stack.
    - Reduced latency of onboard modules by 10 - 25%, up to 94%, through memory optimizations.
    - Customized and optimized Google's protobuf library for up to 2.5x run time speedup in C++.
- Improved reliability of AV onboard system.
  - Developed monitoring on upcoming AV system failures to allow disengagements ahead of time.
  - Enabled automatic issue triage, and reduced hardware & system related P0 issues by 100%.
- Educated and guided engineers on better coding practices through daily work and weekly coding tips.
  - C++ new features and tips, performance optimization tips, tool introductions, etc.

Aug 2012 - Aug 2019 • Research Assistant • Department of Computer Science, **Rice University** • Houston, Texas

- Analyzed and optimized performance of HPC applications.
  - Developed an automated tool to diagnose performance of large-scale parallel applications.
    - Applies novel techniques to sample-based time series data collected with low overhead.
    - Automatically highlights call stacks that are the root causes of performance bottlenecks.
  - Developed a framework that autotunes tensor transposition code for node architectures.

Summer 2015 & 2016 • Summer Intern • **Lawrence Livermore National Laboratory** • Livermore, California

- Developed debugging support for OpenMP and integrated it into the STAT debugging tool for MPI.
- Integrated two performance tools to fit user-annotated knowledge into call stack based performance analysis.

Education & Selective Academic Awards

Aug 2012 – Aug 2019 • **Ph.D.**, Department of Computer Science, **Rice University** • Houston, Texas

- Ph.D. thesis: Automated Diagnosis of Scalability Losses in Parallel Applications

Sep 2008 – Jul 2012 • **B.S.**, Department of Computer Science, **Peking University** • Beijing, China

- Nov 2009 • **Gold Medal** in ***The ACM Asia Programming Contest, Wuhan Site***

Selective Publications

Using sample-based time series data for automated diagnosis of scalability losses in parallel programs, *L. Wei and J. Mellor-Crummey, The 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '20), February 2020.*

Automated Analysis of Time Series Data to Understand Parallel Program Behaviors, *L. Wei and J. Mellor-Crummey, The 32nd ACM International Conference on Supercomputing (ICS '2018), June 2018.*