SI CHEN

4703344900 ◇ sichen.emory@gmail.com ◇ https://www.linkedin.com/in/si-chen-meditates/◇ Suwanee, GA

SUMMARY

Computer Scientist and Engineer specializing in machine learning systems, deep learning models, and system architecture simulation. I focus on applying machine learning techniques to optimize application performance and improve system efficiency. Skilled in workload characterization, performance prediction, and leveraging data to drive intelligent decision-making.

EDUCATION

Emory University *2017 - 2024.12*
Ph.D. in Computer Science
Dissertation title: Efficiently Optimizing HPC Application Design Across a Heterogeneous Hardware Environment

Huazhong University of Science and Technology, China *2004 - 2006*
M.S. in Electrical Engineering

Huazhong University of Science and Technology, China *2000 - 2004*
B.S. in Electrical Engineering

RESEARCH EXPERIENCE

Emory University **Atlanta, GA, USA**
Research Assistant *8/2017 - 12/2024*

Optimized application performance across heterogeneous hardware using HPC simulations and ML

- Achieved a 127× speedup in model training time by developing a cross-architecture HPC prediction system with a Meta-learning model and Gem5 simulator.
- Enhanced the SimPoint architecture simulation acceleration framework using advanced clustering, achieving a 5× speed up in simulation time while maintaining accuracy.
- Utilized feature selection analysis to identify critical hardware performance events with Perf in HPC workloads, reducing data collection time by 95%.

Storage Systems and Workload Characterization

- Developed a gradient-boosting classification model for storage provisioning using time-series-based feature extraction to identify concurrent I/O workloads.
- Designed and implemented workload detection and phase shift prediction pipelines by leveraging FIO trace replay tools for accurate workload modeling.
- Explored workload separation of block I/O trace using blind source separation techniques such as Independent Component Analysis (ICA).

WORK EXPERIENCE

Photon Forces Studio **Suwanee, GA, USA**
Software Development Engineer *1/2025 -*

- Designed and developed a scalable full-stack education web application using React, Node.js, and MongoDB, integrating AI-assisted venue booking features to enhance user experience.
- Built RESTful APIs for customer registration, AI-assisted venue booking, and group notifications.
- Automated form-filling and scheduling through a bot-guided interaction flow, demonstrating an ability to incorporate AI-driven solutions in production systems.

National Center for Atmospheric Research (NCAR) **Boulder, CO, USA**
Research Intern *5/2023 - 12/2023*

- Containerized HPC simulation applications using Docker and Singularity, reducing build time by 60%.
- Implemented CI/CD workflows (GitHub Actions) to automate validation processes, decreasing deployment errors by 40%.

- Deployed containers with diverse MPI/compiler versions across CPU and GPU nodes on supercomputers infrastructures, using Spack for management of software dependencies.

Bytedance **Miami, FL, USA**
Research Intern *5/2021 - 8/2021*

- Developed an AI-powered chatbot using natural language processing (NLP) models to automate error log diagnosis and root cause analysis, reducing support staff workload by 30%.
- Conducted API testing using Postman in the internal deep learning infrastructure platforms.

Netapp **Waltham, MA, USA**
Research Intern *5/2020 - 8/2020*

- Optimized performance headroom predictive metrics for ONTAP data management software using queue theory and the half-latency rule, improving CPU utilization efficiency by 20%.
- Implemented workload merge processes for high availability (HA) failover scenarios, using performance indicators such as IOPS and service time to construct latency—utilization curves.
- Developed workload characterization by statistical analysis of service time distribution (mean and std) in Jupyter NoteBook to improve workload identification precision and curve fitting accuracy.

China Academy of Information and Communications Technology **Beijing, China**
Senior Engineer *07/2006-12/2016*

- Developed the long-term strategy for emergency communication and government communication networks.

TEACHING EXPERIENCE

CS534 Machine Learning (graduate course)	*1/2019 - 5/2019*
CS224 Discrete Structures	*8/2018 - 12/2018*
CS170 Introduction to Computer Science (Java Programming)	*1/2018 - 5/2018*

PUBLICATIONS & PRESENTATIONS

Si Chen, Simon Garcia De Gonzalo, Omar Aaziz, Jeanine Cook, Avani Wildani, *Beyond Guess and Check: Quantifying the Fidelity of Proxy Applications*, SC Workshops PMBS25

Si Chen, Simon Garcia De Gonzalo, Avani Wildani, *MetaCast: Generalizing HPC Application Runtime Prediction*, IPDPS25 (Poster)

Si Chen, Simon Garcia De Gonzalo, Avani Wildani, *SimPoint++: Less Simulation Points*, SC24 Women in HPC Workshop, November 2024

Si Chen, Simon Garcia De Gonzalo, Avani Wildani, *Few-shot HPC application runtime prediction*, Cluster 2023 (Talk + Poster), November 2023

Si Chen, *Research statement*, SySDW23, October 2023

Si Chen, Jianqiao Liu, Avani Wildani, Census*: Counting Interleaved Functional Tenants on Shared Storage*, 36th International Conference on Massive Storage Systems and Technology (MSST 2020), October 2020

Si Chen, Avani Wildani, *Chasing the Signal: Statistically Separating Multi-Tenant I/O Workloads*, workshop on ML for Systems (co-located with NeurIPS 2018), December 2018

SKILLS

Programming Language: Python, C, C++, Java, SQL, Shell scripting
Machine Learning & Data: PyTorch, Tensorflow, scikit-learn
DevOps & Containers: Docker, Singularity, GitHub Actions, Spack
Systems & Tools: Linux, Gem5, Perf, CUDA, HPC clusters, AWS, GCP