Yidi Wang

(408)551-3818 — ywang49@scu.edu — linkedin.com/in/yidi-wang-315649119/

Research Interests — My primary research interests are in the field of real-time embedded and cyber-physical systems. The core objective of my work is to advance the design of energy-efficient, preemptive, and responsive computing systems, particularly when confronted with dynamic timing constraints.

Employment

Santa Clara University — **Santa Clara, CA, USA**
Assistant Professor in Department of Computer Science and Engineering — *Sept 2024 – Present*

University of California, Riverside — **Riverside, CA, USA**
Postdoc in Department of Electrical and Computer Engineering — *Aug 2023 – Jul 2024*

Education

University of California, Riverside — **Riverside, CA, USA**
Ph.D. in Electrical and Computer Engineering — *Sept 2019 – Jun 2023*
- Area of Expertise: Real-time Systems, Embedded Systems, GPUs
- Dissertation: Advancing Real-Time GPU Scheduling: Energy Efficiency and Preemption Strategies
- Advisor: Prof. Hyoseung Kim
- Coursework: Real-Time Embedded Systems, Advanced Operating Systems, Design and Analysis of Algorithms, GPU Architecture and Parallel Programming

University of California, Riverside — **Riverside, CA, USA**
M.S in Electrical and Computer Engineering — *Sept 2018 – Jun 2019*

Huazhong University of Science and Technology — **Wuhan, China**
Bachelor in Electrical Engineering — *Sept 2014 – Jun 2018*

Publications

- Yidi Wang Cong Liu, Daniel Wong, and Hyoseung Kim. GPU Context-Aware Real-Time Scheduling: New Approaches and Improved Analysis. In submission.
- Mohsen Karimi, Yidi Wang, Youngbin Kim, Yoojin Lim, and Hyoseung Kim. CARTOS: A Charging-Aware Real-Time Operating System for Intermittent Batteryless Devices. In submission.
- Yidi Wang Cong Liu, Daniel Wong, and Hyoseung Kim. GCAPS: Analyzable GPU Context-Aware Preemptive Scheduling Approach for Real-Time Tasks. Euromicro Conference on Real-Time Systems (ECRTS), 2024.
- Yidi Wang, Mohsen Karimi, and Hyoseung Kim. Towards Energy-Efficient Real-Time Scheduling of Heterogeneous Multi-GPU Systems. Real-Time Systems Symposium (RTSS), 2022.
- Mohsen Karimi, Yidi Wang, and Hyoseung Kim. An Open-Source Power Monitoring Framework for Real-Time Energy- Aware GPU Scheduling Research. Open Demo Session of IEEE Real-Time Systems Symposium (RTSS@ Work), 2022.
- Mohsen Karimi, Yidi Wang, and Hyoseung Kim. Energy-Adaptive Real-time Sensing for Batteryless Devices. IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), 2022.
- Yidi Wang, Mohsen Karimi, Yecheng Xiang, and Hyoseung Kim. Balancing Energy Efficiency and Real-Time Performance in GPU Scheduling. Real-Time Systems Symposium (RTSS), 2021.
- Yecheng Xiang, Yidi Wang, Hyunjong Choi, Mohsen Karimi and Hyoseung Kim. AegisDNN: Dependable and Timely Execution of DNN Tasks with SGX. Real-Time Systems Symposium (RTSS), 2021.
- Mohsen Karimi, Hyunjong Choi, Yidi Wang, Yecheng Xiang, Hyoseung Kim. Real-Time Task Scheduling on Intermittently Powered Batteryless Devices. In IEEE Internet of Things Journal, 2021.
- Yidi Wang and Hyoseung Kim. Work-in-Progress: Understanding the Effect of Kernel Scheduling on GPU Energy Consumption. In Brief Presentation Session of IEEE Real-Time Systems Symposium (RTSS), 2019.

Reseach Projects

Improving Throughput and Latency for ML Inference Server on AMD GPUs **2023 – Present**

– AMD GPUs are gaining more and more attentions due to its software stack openness. The common issue in GPU resource management is underutilization, and this challenge persists with AMD GPUs. In this research, we aim to enhance the throughput and latency of a PyTorch inference server by leveraging AMD GPU's unique characteristics and addressing their inherent constraints. My primary responsivity in this project is to guide a junior PhD student to walk through the whole process from experiment design to paper writing.

Preemptive Priority-Based Scheduling for Real-time GPU Tasks **2023 – Present**

– This work aims to enable preemptive scheduling mechanisms for real-time GPU workloads at device driver level at contemporary embedded GPU platforms. The main goals of the project include: (1) the development of preemptive strategies for GPU workload with minimum source code modifications, (2) the predictability of GPU timing resource management, and (3) the comprehensive real-time analysis for the strategies.

Energy-Efficient Spatial Multitasking for Heterogeneous GPU Systems **2021 – 2022**

– Due to the lack of SM-level power-gating on GPUs, the spatial multitasking for GPU workloads can unavoidably bring higher energy consumption. This project aims to address this issue in a real-time system. In the project, I developed sBEET, a real-time GPU scheduling framework that dynamically adjusts the degree of spatial multitasking by weighing GPU energy consumption against the system's real-time requirements. The sBEET is further expanded to sBEET-mg focusing on a multi-GPU system.

Real-time Scheduling on Intermittently-Powered Devices **2021 – 2023**

– A significant challenge arises when scheduling tasks on battery-less devices that intermittently derive power from environmental sources. To address this, I have been involved in developing a framework that models energy harvesting and consumption for these devices, ensuring efficient task execution on their processors. My primary contribution in the projects includes the implementation of diverse energy prediction methods, including ML-based strategies, and the formulation of comprehensive schedulability analysis.

Dependable DNN Inference Framework **2021 - 2022**

– This work aims to provide dependable and timely execution of deep neural network (DNN) tasks. The Intel Software Guard Extensions (SGX) enclave provides a secure and isolated execution environment that is tamper-resistant, and the memory and state of the enclave are encrypted. To protect DNN inference from malicious attacks, I have been involved in developing a framework for executing DNN tasks within the secure environment of SGX. My primary contribution to this project includes employing ML-based approaches to obtain a task-level SDC (silent data corruption) probability for a given layer protection configuration.

Teaching Experience

Santa Clara University **Santa Clara, CA, USA**

CSEN20: Introduction to Embedded Systems

– Fall 2024

CSEN283: Operating Systems

– Winter 2025

University of California, Riverside **Riverside, CA, USA**

EE128: Sensing and Actuation for Embedded Systems

– Spring 2023 (Instructor), Spring 2021 (TA), Fall 2020 (TA)

Peer Reviewer

– IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS) - Brief Presentations 2024
– ACM Transactions on Embedded Computing Systems (TECS) 2023 – 2024
– CMTransactions on Cyber-Physical Systems (TCPS) 2023 – 2024
– IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems (TCAD) 2023 – 2024
– IEEE Transactions on Parallel and Distributed Systems (TPDS) 2022 – 2023
– Real-Time Systems Journal 2023
– IEEE Real-Time Systems Symposium (RTSS), SecondaryReviewer 2021

Professional Experience

TuSimple Inc. **San Diego, CA, USA**
Software Development Engineer - Intern *Jun 2022 – Nov 2022*

- Analyzed GPU bottlenecks in self-driving applications and proposed improvements.
- Integrated the improvements into self-driving system to reduce critical path delays.

Wuhan Tianyu Information Industry Co., LTD **Wuhan, China**
Embedded Software Engineer - Intern *Jul 2018 - Aug 2018*

- Migrated essential drivers from a previous embedded system to a new IC card device.
- Worked with the test team to thoroughly test the device, ensuring performance standards and product quality.