Reza Jafarpourmarzouni

Computer Science, School of Engineering, Wayne State University (WSU)
(313)-915-7819 Email LinkedIn Github Website

Summary

Goal-oriented researcher with expertise spanning real-time and embedded systems, scheduling theory, GPU/DNN scheduling, deep learning, and edge-assisted autonomous driving. Strong academic foundation with a Ph.D. in Computer Science from Wayne State University and multiple publications in leading journals and conferences. Skilled in system optimization, resource management, parallel/distributed computing, and AI-driven perception workflows. Experienced in applying C++, Python, and CUDA for real-time scheduling and GPU optimization, and proficient with PyTorch, TensorRT, ONNX, ROS2, and Linux for advancing deep learning and edge-assisted autonomous systems. Passionate about combining theoretical insights with practical systems expertise to create safe, efficient, and intelligent autonomous and real-time platforms.

Technical Skills

- **Programming:** Python, C, C++, CUDA, MATLAB
- **AI/ML Frameworks:** PyTorch, TensorRT, TensorFlow, ONNX, scikit-learn, OpenCV, Hugging Face
- **Real-Time & Embedded Systems:** Real-time scheduling theory (FP, EDF, RM, DM, federated, preemptive, non-preemptive, limited-preemptive), resource management, system optimization, parallel/distributed computing, DAG-task and self-suspending task scheduling, cause-effect chain modeling, tardiness analysis, DNN scheduling
- **Autonomous & Robotics Systems:** ROS/ROS2, simulation platforms (Gazebo, CARLA)
- **Control & Mechanical Systems:** Control theory (linear), dynamic modeling, dynamics fundamentals, MATLAB/Simulink, SolidWorks
- **Model & Workflow Optimization:** Pruning, quantization (FP32/FP16/INT8), workflow acceleration, optimization for edge-assisted systems
- **Development Tools:** VS Code, PyCharm, LaTeX, Git, Docker, Linux, Prompt Engineering, AI-Assisted Coding

Experience/Research Experience

- **Ph.D. Research– Real-Time Systems Group** **WSU | 2022 - Present**
  - Conduct research in real-time and embedded systems, focusing on scheduling theory, GPU/DNN scheduling, and edge-assisted autonomous driving.
  - Designed PreCISE, a GPU scheduling framework enabling limited preemption at DNN layer boundaries, reducing overhead by 90% and improving schedulability by 24.5%.
  - Developed DART, a dependency-aware DAG scheduling framework for autonomous vehicles, integrating node-level offloading and self-suspension analysis; achieved bounded tardiness with improved predictability for DAG tasks in mixed onboard/edge settings.
  - Conducted reaction latency analysis of ROS2 message synchronization (SEAM policy) in edge-assisted autonomous driving; derived safe theoretical bounds for passing and reaction latency, validated through experiments with close alignment to observed maximum latency
  - Conducted real-time inference acceleration research for software-defined vehicles, evaluating TensorRT precision modes (FP32/FP16/INT8) across workflows; achieved substantial throughput gains, with FP16 balancing speed and accuracy and INT8 offering fastest inference
  - Advanced workflow optimization for deep learning in SDVs, integrating pruning and quantization strategies; achieved up to 18× faster inference, 16.5× higher throughput, and 30% reduced GPU/memory usage with minimal accuracy loss
  - Proposed EXcel, an adaptive edge offloading framework for SDVs, achieving up to 7.5× faster inference and 62× lower transmission delays, significantly accelerating both onboard and edge perception workflows.
  - Published in ACM TECS, IEEE IoT-J, and IEEE MOST.
- **Graduate Teaching Assistant** **WSU | 2023 - 2025**
  - Mentored and supported undergraduate and graduate students, fostering understanding of complex computer science topics through one-on-one guidance and group discussions.

- Collaborated with faculty to deliver lectures, labs, and course materials, demonstrating strong teamwork and communication skills.
- Provided constructive feedback on assignments and projects, helping students improve problem-solving and analytical thinking.
- Built leadership and organizational skills by managing office hours, coordinating grading, and assisting students in reaching academic goals.

- **Thomas C. Rumble Fellowship** **WSU | 2022 - 2023**
- **B.S. Mechanical Engineering– Research/Projects** **BNUT | 2016 - 2021**
  - Conducted projects in dynamics, modeling, and control theory, gaining strong foundations in system dynamics and control systems.
  - Hands-on experience with MATLAB/Simulink and SolidWorks for modeling and simulation of control and mechanical systems.

Selected Research Projects & Publications

Reaction Latency Analysis of Message Synchronization in Edge-assisted Autonomous Driving **2025**

R. Jafarpourmarzouni, Sumaiya, R. Li, N. Guan, G. Wang, P. Zhou, Z. Dong

ACM Transactions on Embedded Computing Systems (TECS), Impact Factor: 2.8

- Analyzed synchronization challenges in ROS2 for connected/edge-assisted vehicles and derived safe bounds for reaction/passing latency. Validated the SEAM policy experimentally, showing strong alignment between analytical and observed latencies, ensuring predictable sensor fusion for autonomous driving.

Towards Real-Time and Efficient Perception Workflows in Software-Defined Vehicles **2024**

Sumaiya, R. Jafarpourmarzouni, Y. Luo, S. Lu and Z. Dong

Journal: IEEE Internet of Things Journal (IoTJ), Impact Factor: 10.6

- Addressed throughput, latency, and memory bottlenecks in perception models by integrating pruning and quantization across PyTorch–ONNX–TensorRT workflows. Achieved up to 18× faster inference, 16.5× higher throughput, and 30% lower GPU/memory usage with minimal accuracy loss.

Enhancing Real-time Inference Performance for Time-Critical Software-Defined Vehicles **2024**

Sumaiya, R. Jafarpourmarzouni, S. Lu and Z. Dong

Conference: IEEE International Conference on Mobility: Operations, Services, and Technologies (MOST)

- Optimized object detection for safety-critical SDVs by evaluating TensorRT precision modes (FP32, FP16, INT8). Demonstrated that FP16 offers the best balance between speed and accuracy, while INT8 achieves the fastest inference, guiding deployment of real-time perception models.

DART: Dependency-Aware Real-Time Task Offloading for Edge-Assisted Autonomous Driving

R. Jafarpourmarzouni, Sumaiya, Z. Dong *Submission-Ready*

- Proposed a dependency-aware DAG scheduling framework with node-level offloading and self-suspension analysis. Improved system parallelism and predictability, ensuring bounded response times for safety-critical AV workloads in edge-assisted settings.

PreCISE: GPU-Based Predictable Scheduling for Concurrent DNN Inference in Safety-Critical Environments

R. Jafarpourmarzouni, Sumaiya, Z. Dong *Submission-Ready*

- Introduced a limited-preemptive EDF scheduling framework for DNN tasks by exploiting GPU layer boundaries for preemption. Reduced re-execution overhead by 90% and improved schedulability by 24.5%, enabling predictable inference for multiple concurrent DNNs.

EXcel: Edge-Assisted Real-Time Workload ACceleration for Software-Defined Vehicles

Sumaiya, Y. Luo, R. Jafarpourmarzouni, S. Lu, Z. Dong *Submission-Ready*

- Developed a three-mode offloading framework with transmission-aware and accuracy-aware policies. Combined pruning, quantization, and compression to achieve 4.5× faster onboard inference, 7.5× faster edge inference, and up to 62× transmission delay reduction in real 5G/Wi-Fi environments.

Education

Wayne State University (WSU) **Aug. 2022 – Present**
Ph.D. in Computer Science | ***GPA:*** *3.96/4* *Detroit, MI*

Wayne State University (WSU) **Aug. 2022 – Aug. 2024**
Master of Science in Computer Science | ***GPA:*** *3.96/4* *Detroit, MI*

Babol Noshirvani University of Technology (BNUT) **Sep. 2016 – Feb. 2021**
Bachelor of Science in Mechanical Engineering | ***GPA:*** *16.91/20* *Mazandaran,Iran*

Certificates

C)ISSO: Certified Information Systems Security Officer **Fall 2022**
GPU Programming Specialization **Fall 2025**
Deep Learning Specialization **In Progress**