Collaborative Research/Medium:

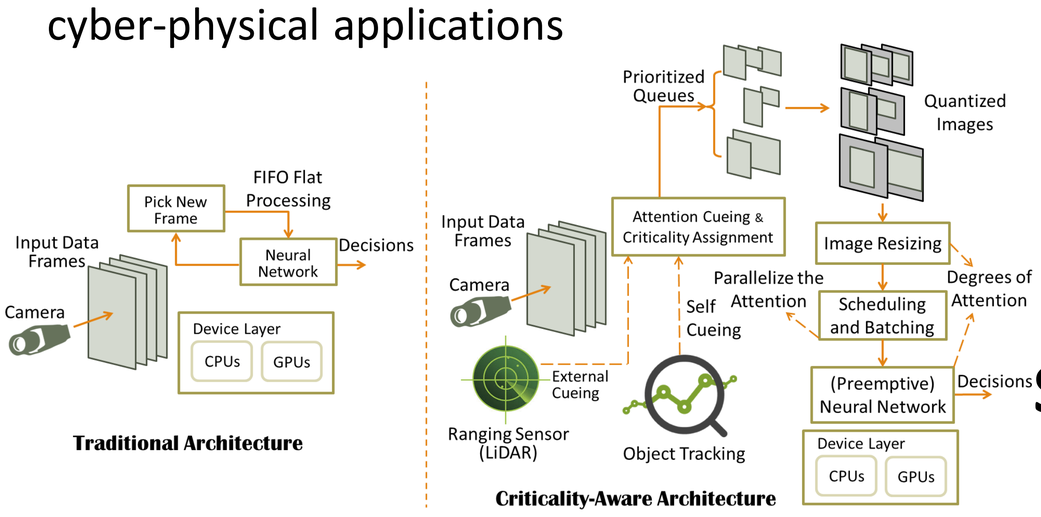
Real-time Criticality-Aware Neural Networks

Tarek Abdelzaher (Lead PI, UIUC), Shuochao Yao (George Mason U.), Heechul Yun (U. Kansas)

https://abdelzaher.cs.illinois.edu/topic-edgeAI.html

Challenge:

Criticality/deadline-unaware processing in current machine intelligence pipelines for cyber-physical applications



Scientific Impact:

Advance the intersection of AI/ML and real-time systems to serve mission-critical CPS applications such as autonomous vehicles and delivery drones

Contributions:

- A general vision for real-time AI [1].
- An architecture for real-time priority-aware machine perception pipelines [2,14].
- Latency models for NN execution on GPUs [3].
- Attention cueing frameworks [4, 5, 6].
- Improved criticality-aware video encoding [5].
- Criticality-aware data input resizing [8].
- Optimizations using contrastive learning [7].
- Deadline-aware 3D point cloud processing [9]
- The DeepPicarMicro testbed for edge AI [10]
- Performance isolation on HMPSoC [11,12,13]
- Performance-assured NN scaling [15,16]
- Generative AI for training perception pipelines [17]
- A book on intelligent edge computing [18]
- Self-supervised pre-training for edge AI applications [19]

Broader Impact:

Societal: Novel solutions will enable more efficient, performant, and robust real-time AI architectures for mission-critical CPS.

Educational:

- Integration with educational modules in "Real-Time Systems" course at UIUC, New "Embedded ML" course at KU, AI summer camp at KU for high-schools.
- A book on Edge AI

Selected Publications:

[1] Tarek Abdelzaher, Sanjoy Baruah, Chris Gill, Eugene Vorobeychik, Ning Zhang, Xuan Zhang, "Research Challenges for Combined Autonomy, AI, and Real-Time Assurance," In Proc. *IEEE CogMI*, December 2021

[2] Shengzhong Liu, Shuochao Yao, Xinzhe Fu, Huajie Shao, Rohan Tabish, Simon Yu, Ayoosh Bansal, Heechul Yun, Lui Sha, Tarek Abdelzaher, "Real-Time Task Scheduling for Machine Perception in Intelligent Cyber-Physical Systems," *IEEE Trans. Computers*, August 2022

[3] Jinyang Li, Runyu Ma, Vikram Sharma Mailthody, Colin Samplawski, Ben Marlin, Songqing Chen, Shuochao Yao, Tarek Abdelzaher, "Towards an Accurate Latency Model for Convolutional Neural Network Layers on GPUs," In Proc. *MILCOM*, November 2021

[4] Shengzhong Liu, Xinzhe Fu, Maggie Wigness, Philip David, Shuochao Yao, Lui Sha, Tarek Abdelzaher, "Self-Cueing Real-Time Attention Scheduling in Criticality-Driven Visual Machine Perception," In Proc. *IEEE RTAS*, Milano, Italy, May 2022

[5] Shengzhong Liu, Tianshi Wang, Jinyang Li, Dachun Sun, Mani Srivastava, and Tarek Abdelzaher, "AdaMask: Enabling Machine-Centric Video Streaming with Adaptive Frame Masking for DNN Inference Offloading," In Proc. *ACM Multimedia*, October 2022

[6] Shengzhong Liu, Tianshi Wang, Hongpeng Guo, Xinzhe Fu, Philip David, Maggie Wigness, Archan Misra and Tarek Abdelzaher, "Multi-View Scheduling of Onboard Live Video Analytics to Minimize Frame Processing Latency," In Proc. *IEEE ICDCS*, July 2022

[7] Dongxin Liu, Peng Wang, Tianshi Wang, Tarek Abdelzaher, "Self-Contrastive Learning Based Semi-Supervised Radio Modulation Classification," In Proc. *MILCOM*, November 2021

[8] Yigong Hu, Shengzhong Liu, Tarek Abdelzaher, Maggie Wigness, and Philip David, "Real-Time Task Scheduling with Image Resizing for Criticality-based Machine Perception," *Journal of Real-time Systems*, Accepted in 2022

[9] Ahmet Soyyigit, Shuochao Yao and Heechul Yun, "Anytime-Lidar: Deadline-aware 3D Object Detection," In Proc. *IEEE RTCSA*, 2022

[10] Michael Bechtel, Qitao Weng and Heechul Yun, "DeepPicarMicro: Applying TinyML to Autonomous Cyber Physical Systems," In Proc. *IEEE RTCSA*, 2022

[11] Michael Garrett Bechtel and Heechul Yun, "Denial-of-Service Attacks on Shared Resources in Intel's Integrated CPU-GPU Platforms," In Proc. *IEEE ISORC*, 2022

[12] Eric Seals, Michael Bechtel, Heechul Yun, "BandWatch: A System-Wide Memory Bandwidth Regulation System for Heterogeneous Multicore," *IEEE RTCSA*, 2023

[13] Michael Garrett Bechtel and Heechul Yun, "Cache Bank-Aware Denial-of-Service Attacks on Multicore ARM Processors," *IEEE RTAS*, 2023

[14] Shengzhong Liu, Shuochao Yao, Xinzhe Fu, Rohan Tabish, Simon Yu, Ayoosh Bansal, Heechul Yun, Lui Sha, and Tarek Abdelzaher, "Taming Algorithmic Priority Inversion in Mission-Critical Perception Pipelines," *Communications of the ACM*, Volume 67, Issue 2, January 2024.

[15] Zhou, J., Li, N., Liu, Y., Yao, S., & Chen, S., "Exploring spherical autoencoder for spherical video content processing," In Proc. *30th ACM International Conference on Multimedia*, 2022.

[16] Leng, Y., Liu, R., Guo, H., Chen, S., & Yao, S., "ScaleFlow: Efficient Deep Vision Pipeline with Closed-Loop Scale-Adaptive Inference," In Proc. *31st ACM International Conference on Multimedia*, 2023.

[17] Tianshi Wang, Jinyang Li, Ruijie Wang, Denizhan Kara, Shengzhong Liu, Davis Wertheimer, Antoni Martin, Raghu Ganti, Mudhakar Srivatsa, and Tarek Abdelzaher, "SudokuSens: Enhancing Deep Learning Robustness for IoT Sensing Applications using a Generative Approach," In Proc. *ACM Sensys*, Istanbul, Turkey, November 2023.

[18] Mudhakar Srivatsa, Tarek Abdelzaher, and Ting He, "Artificial Intelligence for Edge Computing," Springer, 1st edition, December 2023.

[19] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher, "FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space," In Proc. 37th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, Louisiana, December 2023.

Lead PI Contract Info:

Tarek Abdelzaher, Dept. of Computer Science, University of Illinois, Urbana, IL 61801.

zaher@illinois.edu

Award ID# : CNS-2038817, 2038658, 2038923