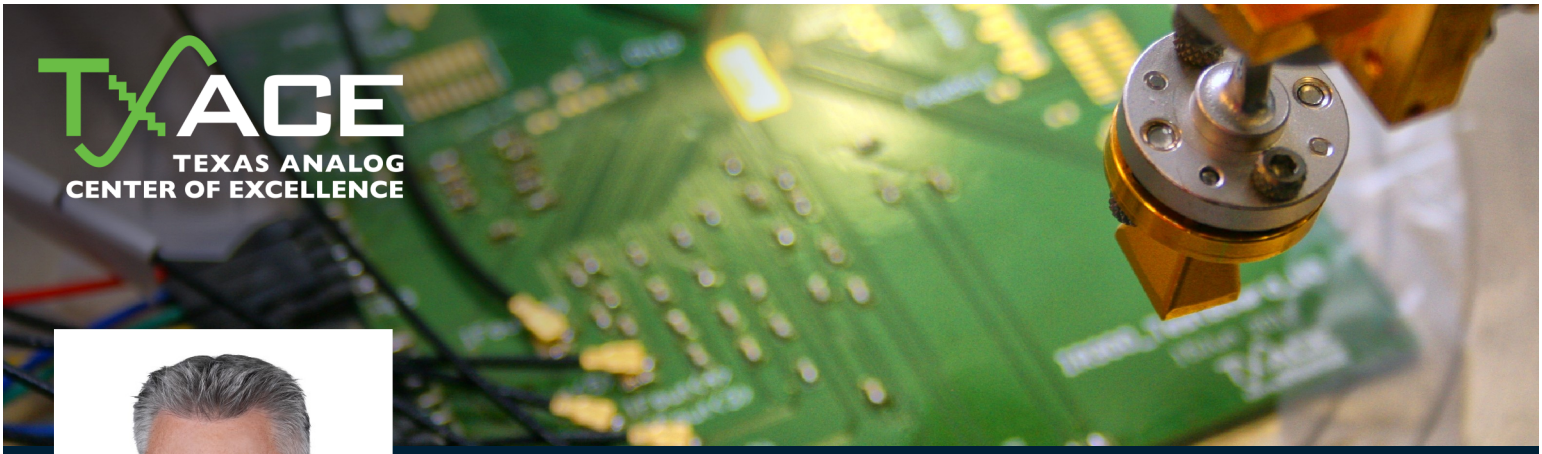
TxACE
TEXAS ANALOG CENTER OF EXCELLENCE



Mixed-Signal Interfaces and Compute Fabrics for tiny Machine Learning (tinyML) Systems

Boris Murmann

Professor of Electrical Engineering
Stanford University

BIO

Boris Murmann is a Professor of Electrical Engineering at Stanford University. He joined Stanford in 2004 after completing his Ph.D. degree in electrical engineering at the University of California, Berkeley in 2003. From 1994 to 1997, he was with Neutron Microelectronics, Germany, where he developed low-power and smart-power ASICs. Since 2004, he has worked as a consultant with numerous Silicon Valley companies. Dr. Murmann's research interests are in mixed-signal integrated circuit design, including sensor interfaces, data conversion, high-speed communication, and embedded machine learning. He was a co-recipient of the Best Student Paper Award at the 2008 and 2021 VLSI Circuits Symposia, as well as a recipient of the Best Invited Paper Award at the 2008 IEEE Custom Integrated Circuits Conference (CICC). He received the 2009 Agilent Early Career Professor Award, the 2012 Friedrich Wilhelm Bessel Research Award by the Humboldt Foundation, and the 2021 SIA-SRC University Researcher Award for lifetime research contributions to the U.S. semiconductor industry. He has served as an Associate Editor of the IEEE Journal of Solid-State Circuits, an AdCom member and Distinguished Lecturer of the IEEE Solid-State Circuits Society (SSCS), the Data Converter Subcommittee Chair and Technical Program Chair of the IEEE International Solid-State Circuits Conference (ISSCC), as well as the Technical Program Co-Chair of the tinyML Research Symposium. He currently serves as the chair of the IEEE SSCS Technical Committee on Open-Source Ecosystem and the General Co-Chair of the 2023 IEEE International Symposium on Circuits and Systems (ISCAS). He is a Fellow of the IEEE.

ABSTRACT

Over the past decade, machine learning algorithms have been deployed in many cloud-centric applications. However, as the application space continues to grow, various algorithms are now embedded "closer to the sensor" and in wearable devices, eliminating the latency, privacy and energy penalties associated with cloud access. In this talk, I will review circuit techniques that can improve the energy efficiency of low-power machine learning inference algorithms at the extreme edge. Specific examples include analog feature extraction for image and audio processing, as well as low-energy compute fabrics for convolutional neural networks. I will present MEDUSA, an end-to-end fully digital input and output stationary in-memory-computing accelerator for commonly used bottleneck layers. Our 28nm CMOS prototype design that is currently being tested achieves 660 nJ/inference on the CIFAR-10 data set.

Friday, March 24, 2023 at 1:00 – 2:00 p.m.
John von Neumann Conference Room
(ECSS 3.910)

UT DALLAS | Erik Jonsson School of Engineering and Computer Science