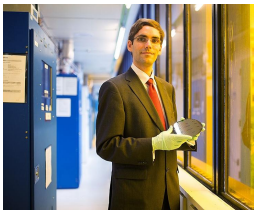
MIT Industrial Liaison Program Faculty Knowledgebase Report

Low Power/Edge Computing

November 5, 2020 11:00 am - 12:10 pm

11:00am	Welcome and Introduction
11:05am	The Extreme Materials Revolution: From Computers in Venus to Synthetic Cells Tomás Palacios Director, MIT Microsystems Technology Laboratories (MTL) Professor, MIT Department of Electrical Engineering and Computer Science (EECS)



Tomás Palacios

Director, MIT Microsystems Technology Laboratories (MTL)
Professor, MIT Department of Electrical Engineering and Computer Science (EECS)

Tomás Palacios is the Director of Microsystems Technology Laboratories (MTL) and is a Professor in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. He received his Ph.D. from the University of California - Santa Barbara in 2006 and his undergraduate degree in Telecommunication Engineering from the Universidad Politécnica de Madrid (Spain). Being a fellow of IEEE his current research focuses on demonstrating new electronic devices and applications for novel semiconductor materials such as graphene and gallium nitride. Tomás is passionate about making an impact on modern society in Energy, Engineering, Nanoscale, Physics, Semiconductors, Nanotechnology, and Climate Change. His work has been recognized with multiple awards, including the Presidential Early Career Award for Scientists and Engineers, the 2012 and 2019 IEEE George Smith Awards, and the NSF, ONR, and DARPA Young Faculty Awards, among many others. Prof. Palacios is the founder and director of the MIT MTL Center for Graphene Devices and 2D Systems, as well as the Chief Advisor and co-founder of Finwave Semiconductor, Inc. From 2023, Tomas serves as Associate Director of the SUPeRior Energy-efficient Materials and Devices (SUPREME) center, one of the seven 2023 JUMP 2.0 programs sponsored by Semiconductor Research Corporation.

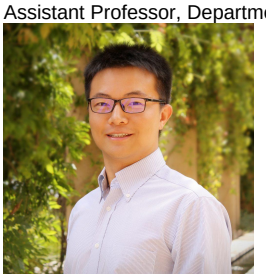
View full bio

The end of traditional transistor scaling brings unprecedented new opportunities to semiconductor devices and electronics. In this new era, heterogeneous integration of new materials becomes key in order to add new functionality and value to electronic chips. This talk will review some examples of these new opportunities, including 1. Gallium Nitride vertical power transistors and CMOS logic for a much more efficient electric grid; 2. One-layer-thick molybdenum disulfide wi-fi energy harvesters to enable ubiquitous electronics; 3. High temperature CMOS technology to power future missions to Venus; and 4. A new generation of cell-sized autonomous electronic microsystems to revolutionize environmental monitoring and healthcare. The seminar will conclude with a reflection on how the democratization of heterogeneous integration and the unique properties of extreme materials will transform our society just as Moore’s law has done for the last 50 years.

11:50am

MCUNet: TinyNAS and TinyEngine for Efficient Deep Learning on Microcontrollers

Song Han
Assistant Professor, Department of Electrical Engineering and Computer Science, MIT EECS



Song Han
Assistant Professor, Department of Electrical Engineering and Computer Science
MIT EECS

Song Han is an assistant professor in MIT's Department of Electrical Engineering and Computer Science. His research focuses on efficient deep learning computing. He has proposed “deep compression” as a way to reduce neural network size by an order of magnitude, and the hardware implementation “efficient inference engine” that first exploited model compression and weight sparsity in deep learning accelerators. He has received best paper awards at the International Conference on Learning Representations and Field-Programmable Gate Arrays symposium. He is also a recipient of an NSF Career Award and MIT Tech Review's 35 Innovators Under 35 award. Many of his pruning, compression, and acceleration techniques have been integrated into commercial artificial intelligence chips. He earned a PhD in electrical engineering from Stanford University.

Machine learning on tiny IoT devices based on microcontroller units (MCU) is appealing but challenging: the memory of microcontrollers is 2-3 orders of magnitude less than mobile phones, not to mention GPUs. I will introduce key technologies for neural network optimization on IoT devices, including model compression (pruning, quantization), neural architecture search, and compiler/runtime optimizations. Based on that, we propose MCUNet , a framework that jointly designs the efficient neural architecture (TinyNAS) and the lightweight inference engine (TinyEngine). MCUNet automatically designs perfectly matched neural architecture and the inference library on MCU. MCUNet enables ImageNet-scale inference on microcontrollers that has only 1MB of FLASH and 320KB SRAM. It achieves significant speedup compared to existing MCU libraries: TF-Lite Micro, CMSIS-NN, and MicroTVM. Our study suggests that the era of tiny machine learning on IoT devices has arrived.