

## Research Overview:

My research interests lie in the intersection of electronic design automation (EDA), computing systems, and computer architecture, with a focus on improving the **energy efficiency** of computing systems and designing **intelligent** techniques to help automate computing system management. Differentiating themes of my research are: (1) synthesizing design and runtime telemetry across the different hardware and software layers in a system to adapt to varying application, system, or environmental conditions, (2) focusing on the complex interplay among several important metrics, such as performance, energy, and temperature, while most work in the area considers one or two of these parameters, and (3) applying machine intelligence towards the design and operation of computing systems to push performance, efficiency, and resilience dramatically beyond the state-of-the-art. I have received over \$4M of funding from federal grants and over \$800K of research gifts from industry and other sources to support my research. My group has collaborated with a number of companies (e.g., IBM, AMD, Oracle, Intel), national and international research labs (e.g., Sandia Labs, CEA-LETI in France), centers, and university research groups (e.g., at MIT, Brown, EPFL, and others), and we will continue to pursue and foster such collaborations.

My research group currently pursues computer engineering research in two main fronts: (1) *designing future energy-efficient computing systems* and (2) *designing analytics and optimization methods for large-scale computing systems (such as for HPC and cloud)*. As part of the first thread, we are particularly focused on understanding how new device and integration technologies impact future computing system design. Specifically, we have explored 3D/2.5D/monolithic-3D integration, silicon-photonic on-chip interconnects, and on-chip integration of emerging cooling technologies—all of which are now separately funded projects, spun off from my NSF CAREER project. My team's research investigates how to model such technologies at the architecture and system level. We demonstrate benefits and challenges of new technologies with a full system view and while emulating different applications running on that system (i.e., in contrast to solely focusing on device or circuit-level optimization), and in this way, devise design and runtime strategies that enable achieving the most out of future systems, without leaving substantial efficiency or performance benefits on the table. My work on energy-efficient computing has been recognized with the [Ernest Kuh Early Career Award](#). Recent key publications include: [DATE'19](#) and [TCAD'17](#) (silicon photonics); [TCAD'20](#) and [DATE'20-best paper nominee](#) (designing with emerging cooling methods); [TCAD'20](#) (2.5D design optimization); [TCAD'19](#) (mobile system efficiency).

During my time at BU and especially over the last 5 years, I have constructed a brand-new research front in my group on large-scale systems analytics and optimization. This second main front addresses the problems resulting from the rapidly increasing size and complexity of computing systems that serve the cloud and HPC. Our research has focused on designing frameworks that effectively synthesize monitoring telemetry (or other relevant system and software information) from tens of thousands of virtual or physical nodes, train for known problematic software instances or performance problems, and then at near-real-time diagnose or identify performance bugs or vulnerable code. This automation approach is disruptive as much of the system management relies on heavy expert involvement. Recent key publications include: [TPDS'19](#), [EuroPar'18](#), and [ISC-HPC'17](#) (HPC performance anomaly diagnosis); [TCC'20](#), [SoCC'19](#), and [BigData'16](#) (cloud analytics). My team will continue to make strides in this front, especially in designing *scalable and explainable* AI frameworks for cloud and HPC, as much of the “black-box” machine learning methods are not easily applied to real-world systems due to challenges with the lack of understanding and trust in

automated decisions. We started a project on this topic with Sandia Labs recently. Automating system management via machine learning based approaches has a number of direct impacts in performance, efficiency, resilience, and security, and we will continue to demonstrate quantifiable impact.

An important component of my research has been demonstrating impact in real-world scenarios. I have always emphasized demonstrating results in prototypes, demos, artifacts as much as possible. In the last few years, we have released several [tools](#) (e.g., thermal modeling tools for 3D chips) and software artifacts (e.g., [Taxonomist](#) best artifact award at EuroPar'18), and demonstrated our tools in several conferences (e.g., [Praxi software discovery tool](#) demonstrated at the Middleware'19 conference). We will continue this trend of open-sourcing software and data to speed up innovation and reduction to practice.

Another key component of my research has been to identify problems in “intersection” areas, such as at the intersection of computer systems and machine learning, or at the intersection of architecture and emerging devices. One such project I have constructed at BU has been on integrating high-power-consuming data centers into emerging power programs in the smart grid. We published important papers in this direction (e.g., [TOMPECS'19](#), [ICCAD'13](#)) and now we have arrived at a point of real-world impact, where we are able to design data center management and power market participation policies with guarantees on quality-of-service delivered to users ([eEnergy'19](#)). Our project will continue to bridge the theoretical gains we have demonstrated so far with practical scenarios in real systems—an often overlooked aspect in this line of research.