Title: Sequential abstractions of data-parallelism in hardware
Abstract:

Semi-programmable hardware accelerators are ubiquitous in modern systems-on-chip but verifying correctness properties about the interaction between hardware accelerators and software is very challenging. In particular, many accelerators for machine learning use data-parallelism in hardware, and existing modeling and verification techniques would typically need to perform expensive reasoning about the interleaved updates to states of the system by the parallel hardware components.

In this talk, I present my work on two sequential abstractions of data-parallelism in hardware, "SIMD instructions" and "n-way parallelism", which I use to extend an existing specification model for hardware called an ILA (instruction level abstraction). An ILA provides architecture-level specifications for hardware accelerators in much the same way that an ISA (instruction set architecture) provides specifications for processors. Essentially, an ILA instruction corresponds to certain inputs at the interface of the accelerator and captures the updates to architectural state that are performed by the accelerator in response to those inputs.

For these two abstractions, I show how each abstraction allows an ILA instruction to capture the parallel updates due to data-parallelism in hardware, while still maintaining a sequential state transition semantics. This avoids expensive reasoning in verification due to interleaving of the parallel updates. Next, I show a further use of these abstractions, defining a "SIMD transformation" that can take a low-level model of a data-parallel accelerator and produce a higher-level model that is guaranteed to be equivalent to the low-level model. This has applications in automated code generation for accelerator hardware. Finally, I discuss a proof-of-concept implementation of these abstractions in ILAng, a platform for modeling and verification of software/hardware systems using ILA models. I will demonstrate the use of these abstractions on a real-world accelerator designed for machine learning and describe the future directions the work can be extended.