Spatial Computing Architectures for Deep Learning Workloads

Abstract: To meet the stringent energy-efficiency and performance needs of deep learning workloads, minimizing data movement is crucial. This trend plays to the strengths of spatial architectures, including not only FPGAs but also processor arrays with mesh interconnects like the Cerebras Wafer Scale Engine and AMD AI Engine arrays. Spatial architectures also have characteristics that make them particularly well-suited to More-than-Moore integration that stacks multiple dice, connects them with interposers, or leverages wafer-scale integration. To explore this broad architecture space, we need to not only update FPGA CAD tools, but to also create entirely new tool flows for spatial processor arrays. In this talk, I will discuss the relative strengths of these spatial platforms vs. GPUs across a range of deep learning applications, outline how the different platforms are exploiting More-than-Moore integration, and give an overview of the tools we are creating to help enable and explore 3D FPGAs, new embedded FPGAs, and spatial processor arrays.

Biography: Dr. Betz is a Professor of Electrical and Computer Engineering at the University of Toronto and a Distinguished Visiting Professor at Cerebras Systems. His research interests cover VLSI design and computer architecture (particularly FPGA and spatial processor array architecture), Computer-Aided Design algorithms, and accelerating deep learning inference.

He is the original developer of the VPR FPGA architecture exploration tool and the leader of the open-source Verilog-to-Routing (VTR) project. He co-founded Right Track CAD to commercialize his research; after its acquisition he spent 11 years at Altera, ultimately as Senior Directory of Software Engineering. He is an architect of the Quartus CAD system and the first five generations of the Altera Stratix and Cyclone families.

Dr. Betz holds 102 US patents, and has received 18 best or most significant paper awards from top conferences and journals in the field. He is Editor-in-Chief of the ACM Transactions on Reconfigurable Technology and Systems journal and serves on the program committees of several conferences in the CAD and programmable silicon areas. He is an IEEE Fellow, an ACM Fellow, a Fellow of the National Academy of Inventors, a Fellow of the Engineering Institute of Canada, and a Faculty Affiliate of the Vector Institute for Artificial Intelligence.