ARIZONA

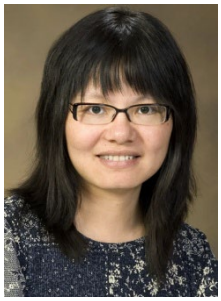
SEMICONDUCTORS & MICROELECTRONICS
DESIGN FOR LARGE-SCALE AI INFERENCE

University of Arizona researchers investigate new analog near-threshold designs that can achieve circuit and system-level robustness with small power consumption and small area requirements for computation.

Built upon the scientific disciplines of signal processing and artificial intelligence, AI is emerging as an essential technology platform for vision, decision-making, machine learning, human-computer interface, and many more applications. A critical component of AI is inference from incomplete or erroneous data. Inference can be found in many modern applications including video analytics, compressive sensing, and big data (e.g., drug discovery, stock prediction, bioinformatics, etc.). While it can be cast rigorously in terms of underlying probabilities and benefits from a strong theoretical foundation, inference may also lead to high computational complexity, for example, NP-hard. To reduce computational cost, powerful modern approaches seek approximate solutions via message passing (MP) on graphical models. Although MP is based on simple computational primitives, overall complexity can become intractable for large-scale graphs. To implement a 1 million node graph, a modern digital ASIC implementation for MP will consume 1.5 GW and require 5776 m^2 (roughly 3x the increased area and power consumption expected for CPU-based implementations).

We use the near-threshold design to investigate Ising problems that are believed to be solvable only using quantum designs, not semiconductor technologies. An essential aspect of this work is evaluation of the robustness of near-threshold designs to implement imperfections. In addition to traditional problems in graphical inference, our recent results suggest that MP can be applied to generalized Ising problems (i.e., finding a collection of spins that minimize the Ising Hamiltonian), which is NP-hard. We implemented the proposed graphical inference approach using TSMC 65nm. Early results suggest that MP computation convergence with less than 10% message errors is possible, considering imperfections such as process variations, transistor layout mismatch, and 1/*f* noise. Projecting our results to a 14 nm technology node, we demonstrated that the near-threshold computation has substantial benefits: 70kW power consumption for 1 million nodes, with 2 cm^2 per node as area consumption.

Janet Roveda, PhD | Professor
Electrical & Computer Engineering | meilingw@arizona.edu



Janet's research interests include robust VLSI circuit design, biomedical instrumentation design, smart grid, VLSI circuit modeling/design and analysis, and low power multi-core system design. Her groundbreaking research has earned her prestigious honors, including the PECASE award, and an AIMBE Fellowship. She has over 150 scholarly publications.

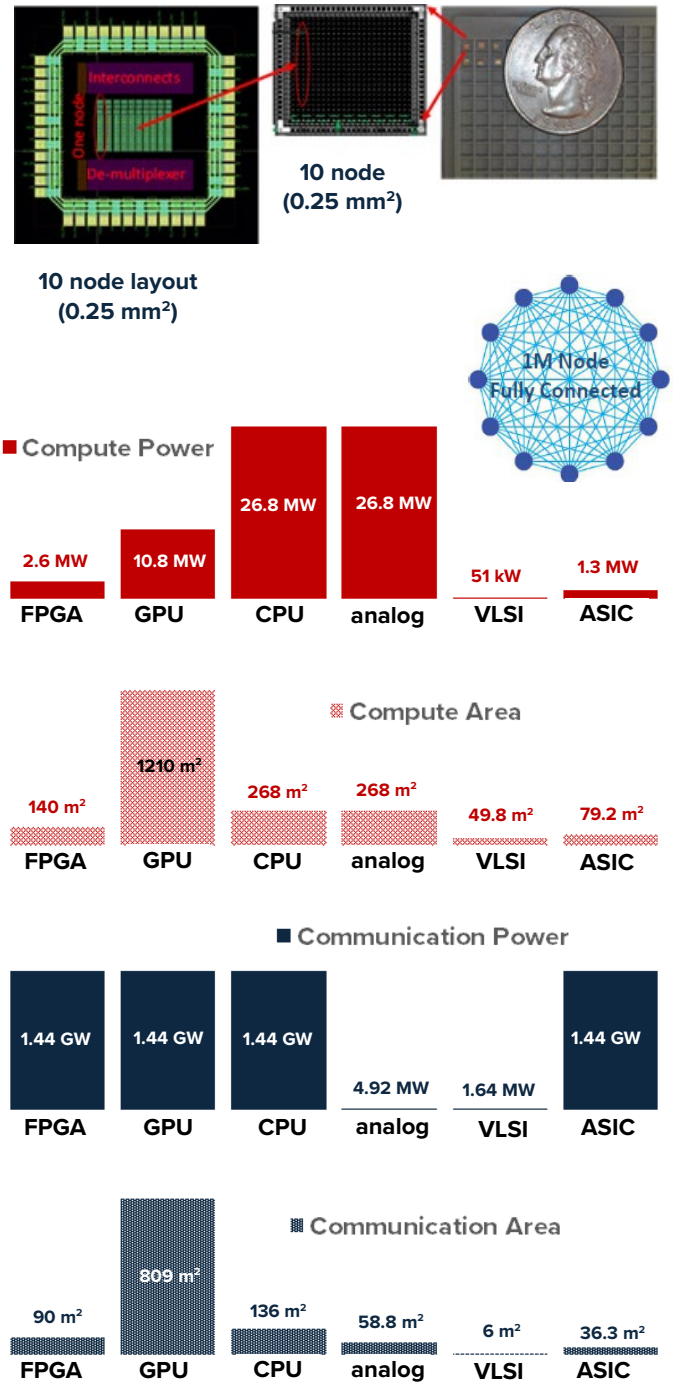


Figure: Our subthreshold very large-scale integration (VLSI) circuit designs have substantial benefits. Projecting our results for 1 million fully-connected nodes dramatically reduces power and area consumption based on message passing on graphical models. Compare with traditional field-programmable gate arrays (FGPA), graphical and central processing units (GPUs, CPUs), analog circuits and modern application-specific integrated circuits (ASICs).