Building the Next Generation of Chips with Generative AI

Sherief Reda
School of Engineering and Computer Science Dept, Brown University

The semiconductor industry faces a critical challenge: the cost and complexity of designing advanced System-on-Chips (SoCs) are increasing exponentially, creating an unsustainable productivity gap. As chip designs become more intricate, scaling up by adding more engineers or tools is no longer a viable solution. Generative AI (GenAI) presents a transformative approach to this problem by decoupling design complexity from manual human effort, dramatically increasing the productivity of individual engineers, and finding superior solutions to computationally intensive problems in chip design. This tutorial will provide an overview of the major research themes and latest results in the use of GenAI for digital chip design.

A primary application of GenAI is the generation of synthesizable Verilog code from natural language descriptions using Large Language Models (LLMs). A critical challenge is achieving high synthesis quality. This is measured not just by whether the code works correctly but by how efficient the resulting circuit is in terms of Power, Performance, and Area (PPA). We provide a large-scale evaluation of LLM-generated Verilog quality using a wide range of models and designs. We show how techniques like Chain-of-Thought (CoT) prompting and fine tuning are used to guide the model's reasoning, compelling it to break down problems logically, much like a human designer would.

GenAI is also being deployed to address a significant bottleneck in modern chip design: verification. This process, which can consume up to 60% of a designer's time, is being automated as LLMs learn to generate complex verification environments, testbenches, and formal properties (i.e., SystemVerilog Assertions) directly from design specifications. We will describe state-of-the-art methods for verification and show how to design LLM agents that can create testbenches in order to maximize functional coverage.

Beyond code generation, GenAI is revolutionizing how engineers interact with vast and complex design data. We describe how Retrieval-Augmented Generation, frameworks like ChipXplore act as intelligent assistants. They allow engineers to ask natural language questions about massive datasets like Process Design Kits and design databases. The LLM-based agentic flow translates these questions into database queries, retrieving critical information faster and with fewer errors than manual methods, thereby dramatically improving productivity and accuracy.

The tutorial will also describe how GenAI is used to create predictive tools and autonomous agents to automate the entire chip design flow. LLMs can analyze RTL code directly to estimate a circuit's PPA metrics before synthesis, providing rapid feedback and enabling early design space exploration. We will show how GenAI techniques can create autonomous multi-agent systems that can manage the entire design workflow, from RTL to the final GDSII layout. In this paradigm, an LLM acts as a high-level reasoning engine, planning tasks, generating scripts for design automation tools, and executing the design flow. This agentic flow can continuously evaluate the results of each step and iteratively refine the design until all objectives are met.

The tutorial will conclude with open research problems in the field, including the scarcity of high-quality public design data, the risk of hallucinations, the risk of IP violations, and the need to create multi-modal autonomous AI-driven agentic flows that can integrate textual specifications, structural code, and visual layout data to holistically optimize chip designs.

Biography:

Sherief Reda is a Full Professor at the School of Engineering and the Department of Computer Science, Brown University. He received his BSc (with Honors) and MSc in Electrical & Computer Engineering from Ain Shams University, Cairo, Egypt. He joined the School of Engineering at Brown University in 2006 after receiving his PhD in Computer Science & Engineering from the University of California, San Diego. His research interests are in the areas of energy-efficient computation, design automation, and embedded systems. He is an IEEE Fellow for his contributions to energy-efficient and approximate computing. Professor Reda has co-authored or edited two books and has over 150 articles in leading conferences and journals. Professor Reda has received several best paper awards. He has been a PI or co-PI on more than $22M in funded projects from federal agencies and industry corporations. Professor Reda is a holder of five US patents, some of which have been licensed and commercialized. He served as a technical program committee member for many IEEE/ACM conferences in his research area, and as an associate editor for Integration, the VLSI Journal (Elsevier) and IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD). Besides his academic work, he was a Principal Scientist at Amazon between 2021 and 2023, and he also serves as an expert witness in patent litigation lawsuits. He is currently an Amazon Scholar. He is a recipient of an NSF CAREER Award and is both an IEEE and an AAIA Fellow.

Related Prior Publications

- M. Abdelatty, J. Rosenstein, and S. Reda, ChipXplore: Natural Language Exploration of Hardware Designs and Libraries, IEEE International Conference on LLM-Aided Design (LAD), **Best Paper Award**, 2025
- M. Abdelatty, J. Ma, S. Reda, MetRex: A Benchmark for Verilog Code Metric Reasoning Using LLMs, IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), 2025
- M. Abdelatty M. Nouh, S. Reda, PLUTO: A Benchmark For Evaluating Efficiency OF LLM-Generated hardware code, under review in International Conference on Learning Representations (ICLR), 2025
- A. Agiza, R. Roy, T. Ene, S. Godil, S. Reda and B. Catanzaro GraPhSyM: Graph Physical Synthesis Model, Proceedings of the 42nd International Conference on Computer-Aided Design (ICCAD), 2023
- A. Hosny and S. Reda, Characterizing and Optimizing EDA Flows for the Cloud, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2022.
- T. Ajayi, V. A. Chhabria, M. Fogaca, S. Hashemi, C. Holehouse, A. Hosny, A. B. Kahng, M. Kim, J. Lee, U. Mallappa, M. Neseem, G. Pradipta, S. Reda, M. Saligane, S. S. Sapatnekar, C. Sechen, M. Shalan, W. Swartz, L. Wang, Z. Wang, M. Woo and B. Xu, Toward an Open-Source Digital Flow: First Learnings from the OpenROAD Project, IEEE/ACM Design Automation Conference, 2019