Master's Project

Optimal routing for 2D Mesh-based Analog Compute-In-Memory Accelerator Architecture

The explosive application of **deep neural networks** to every conceivable domain has led to a consequent interest in acceleration architectures for performing efficient inference of very large networks. While GPUs remain the leading device of choice for DNN implementation, **more exotic accelerators** including dedicated ASIC designs and in-memory based designs have also been proposed.

In this regard, IBM has designed and demonstrated a **2D mesh-based Analog Compute-In-Memory accelerator** for DNN inference, pictured in Figure 1. The accelerator consists of a mesh of nodes, containing a mix of analog-compute tiles that perform efficient Vector-Matrix Multiplications (VMMs), and digital tiles to handle intermediate digital operations. This accelerator is capable of performing inference at **significantly lower latencies** and with better **power efficiency** in comparison to SotA works.

One challenge of such mesh-based architectures is that of **routing**, namely, how are network layers and operations mapped onto the mesh in order to reduce latencies and avoid mesh contention. Such challenges are analogous to those found in fields such as digital circuit and FPGA synthesis, where such algorithms have been studies for decades and a wide variety of open-source libraries are already available.

We are looking for enthusiastic Master's students interested in **gaining practical experience** in exploring SotA routing algorithms and applying them to cutting edge analog in-memory computing architectures for DNNs.

The position is expected to take place at **IBM's Zurich labs** for a duration of **a semester**.

Requirements

- Outstanding programming skills (C/C++ and/or Python
- Independent learning/working abilities
- Interest in FPGAs and/or digital synthesis flows
- Strong work ethic

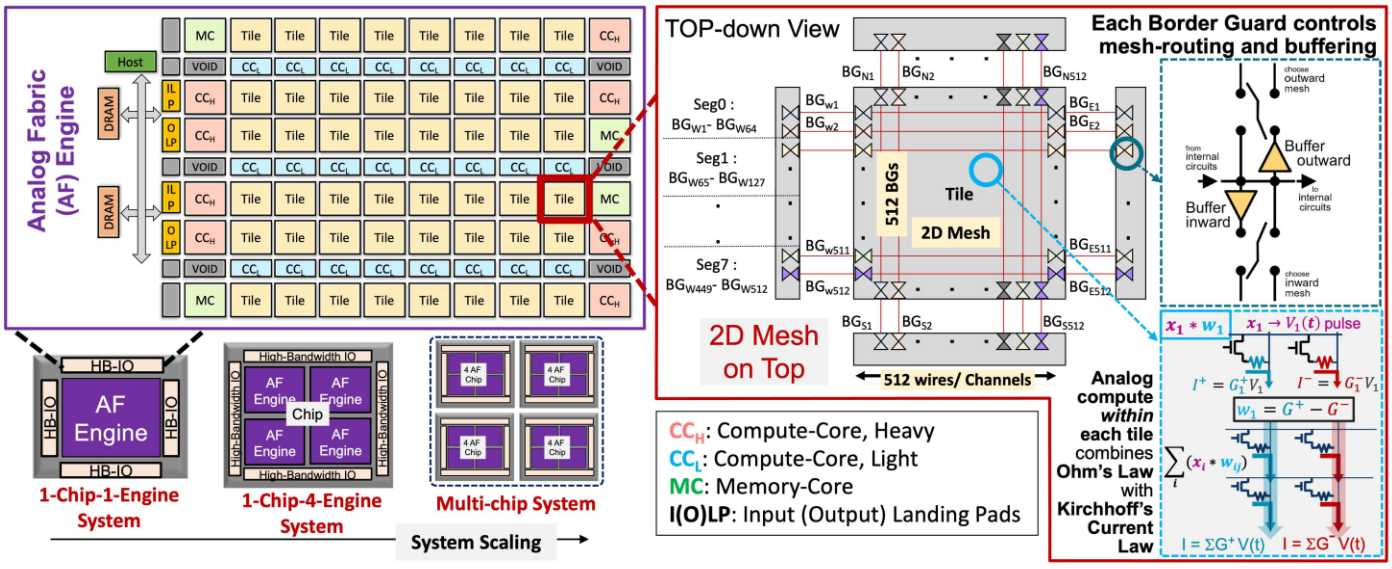

Figure 1: 2D Mesh-based Analog Compute-In-Memory Accelerator Architecture

For background and past works, see:

- Analog-In Memory Processing
  - A Heterogeneous and Programmable Compute-In-Memory Accelerator Architecture for Analog-AI Using Dense 2-D Mesh | IEEE Journals & Magazine | IEEE Xplore
- FPGA Routing
  - Tutorial on FPGA Routing
- Open source PnR libraries
  - VPR — Verilog-to-Routing
  - RapidWright

If you are interested, please contact **Dr. William Simon:**

william.simon1@ibm.com