EE290C Project Proposal: Performance and Energy Characterization of Gemmini, NVDLA, and BOOM Using State-Sampling from Firesim FPGA Simulation

Vighnesh Iyer, Billy Chau

1 Background and Motivation

The architecture and VLSI literature has produced a litany of dedicated ML accelerators over the last 3 years. Most of these accelerators exploited unique dataflows, weight/activation sparsity, integer arithmetic, novel circuit techniques, or many other techniques to achieve a claimed perf/watt advantage over prior work.

However, architecture papers typically measure the energy consumption of the proposed accelerator against a baseline using a (Python/C++) model of their accelerator, and the energy numbers may suffer from poor accuracy. Some papers use HLS to generate RTL implementations of different architectures but the energy calculation from a synthesized netlist and RTL simulation is limited by the speed of simulation, and thus only small networks can be evaluated at the RTL-power-estimation-level.

On the other hand, chip/VLSI papers measure the energy consumption of their accelerator empirically using the taped out chip, but do not have a reference accelerator implementation taped-out to compare against. Furthermore, once a chip is taped out, the power consumption cannot be measured for individual parts of the accelerator (at module-granularity) and cycle-by-cycle energy numbers cannot be extracted.

2 Prior Work

3 Proposed Extension Over Prior Work

4 Project Infrastructure

5 Project Timeline

1. **Checkpoint 1 (4/10)**:
2. **Checkpoint 2 (4/24)**:
3. **Final Report (5/8)**: