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

Strober[3] is a framework which can generate fast and cycle-accurate sample-based energy simulator using Chisel for any arbitrary RTL running on FPGA. It has achieved less than 5 percents error with 99.9 percents confidence against commercial CAD tools with more than two orders of magnitude speedup over existing microarchitectural simulators and four orders of magnitude speedup over commercial Verilog simulators.

State Snapshotting using Scan Chain - As detailed information of simulation is required for an accurate power model, a basic scan chain is embedded into the design to capture a replayable RTL snapshot with both register and SRAM values. Using the hardware construction language Chisel, all custom transforms and FAME1 transform can be easily mapped and wrapped around the RTL design.

Sample-based Energy Simulation - Based on the central limit theorem of statistics, a representative estimator of the power model can be generated given random sampling and enough snapshot samples. However, knowing the length of the program's execution is impossible, so the reservoir sampling technique is used to address this problem.

Fine-grained level Simulation and Checking - Independent replayable snapshots are replayed in

parallel in industrial gate-level simulators such as VCS for accurate power analysis. To ensure the correctness of the execution, the output values of the design are compared with the output traces.

Simmani[2]

DESSERT[1]

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)**:

References

[1] Donggyu Kim et al. "DESSERT: Debugging RTL Effectively with State Snapshotting for Error Replays across Trillions of Cycles". In: *2018 28th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, Aug. 2018.

[2] Donggyu Kim et al. "Simmani: Runtime Power Modeling for Arbitrary RTL with Automatic Signal Selection". In: *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2019, Columbus, OH, USA, October 12-16, 2019*. ACM, 2019, pp. 1050–1062.

[3] Donggyu Kim et al. "Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL". In: *43rd ACM/IEEE Annual International Symposium on Computer Architecture, ISCA 2016, Seoul, South Korea, June 18-22, 2016*. IEEE Computer Society, 2016, pp. 128–139.