INTELLIGENT MACHINE ENGINEERING CONSORTIUM
Machine Learning, Reasoning, and Understanding from Cloud to Edge

The University of Texas at Austin

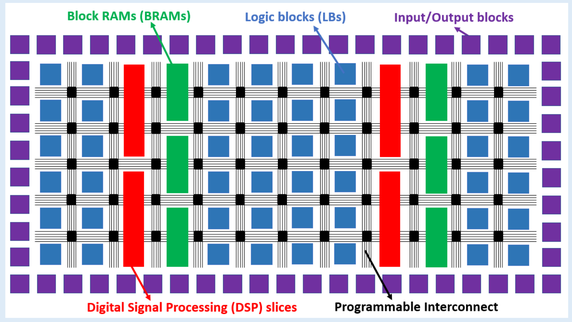
Efficient Reconfigurable Substrates for Deep Learning

Aman Arora, Lizy K. John (University of Texas at Austin)

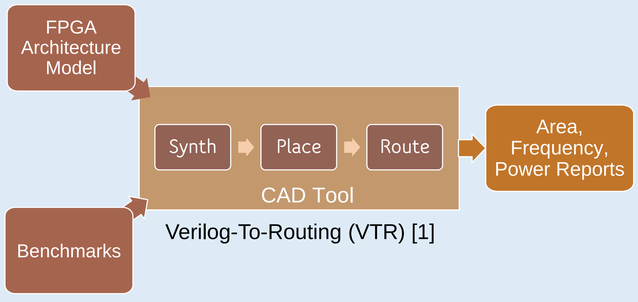
Introduction

- FPGAs (Field Programmable Gate Arrays) are very well-suited to the fast-changing needs of Deep Learning (DL) applications.
- FPGA advantages:
  - High parallelism
  - Customization/Reconfigurability
  - Low latency and high energy efficiency
  - Low NRE costs
- FPGAs trail behind ASICs (Application Specific Integrated Circuits) in terms of performance and efficiency.
- Proposal: **Domain-Optimized FPGAs**

FPGA Architecture Exploration



FPGA primarily consist of blocks such as Logic Blocks, Block RAMs and DSP Slices, immersed in a sea of programmable routing/interconnect.



Verilog-To-Routing (VTR) [1]

FPGA architecture exploration involves taking a set of benchmarks through a CAD tool for various architecture options, and comparing the resulting metrics

Koios Benchmarks

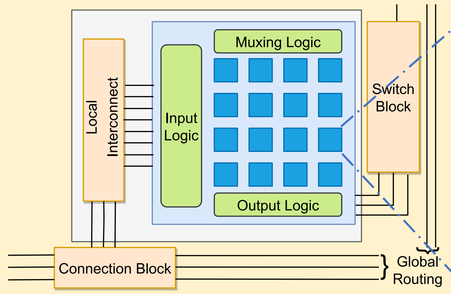
https://tinyurl.com/vtrkoios

- Representative benchmarks are essential for architecture optimization. Existing FPGA benchmark suites do not cater to DL applications.
- We created Koios [4], a DL-specific benchmark suite for FPGA architecture and CAD research. It is open-source and works with VTR.
- It contains 40 benchmarks covering a diverse space within DL and includes synthetic/proxy benchmarks.

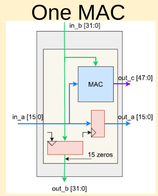
Tensor Slices

- Tensor/matrix operations are at the heart of Deep Learning
- Matrix multiplier using traditional compute blocks (Logic Blocks and DSP Slices) is inefficient (~4x slower and ~10x larger than an ASIC). Can we perform matrix multiplication on an FPGA more efficiently?
- We propose adding new blocks called Tensor Slices [2] to an FPGA. These are compute blocks containing a 2D array of processing elements that can efficiently perform tensor/matrix operations.

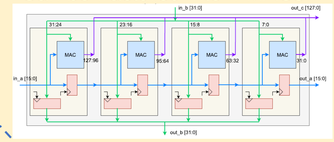
High Level Diagram



One MAC



Can be fractured into 4 MACs



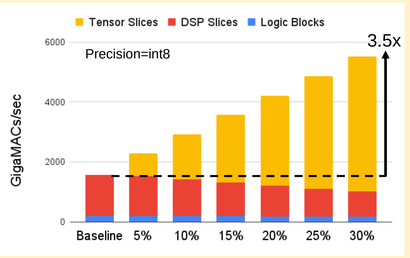
Features

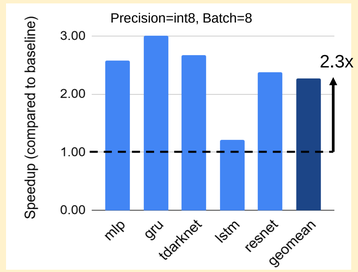
- Architecture: Systolic
- Sizes: 4x4, 8x8
- Precisions: int8, int16, fp16, bf16
- Layout: Columnar
- Operations: MMM, MVM, Eltwise

Pros and Cons

Increase compute throughput of the FPGA
Increase frequency; speed up DL applications
Reduce routing wire usage, hence power consumption
Coarse grained. Faster compilation
Need to update tools, provide libraries, etc.
Less generic/flexible than a typical FPGA.

Results

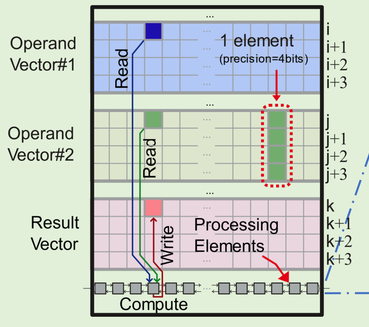


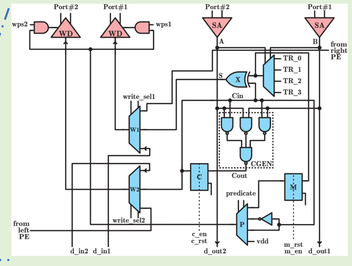


CoMeFa RAMs

- Hundreds of megabits of on-chip memory in form of BRAMs
- Provide high bandwidth to compute units via the programmable routing/interconnect. Can we bring the data even closer to compute?
- We propose CoMeFa RAMs [3] which are memory blocks that have computation capabilities. Bit-serial processing elements are added inside the RAM. A transposed data layout is used for computation.

High Level Diagram





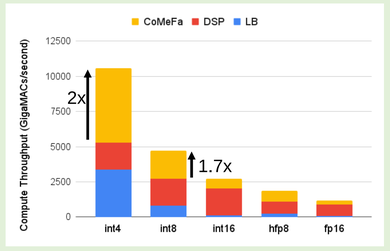
Features

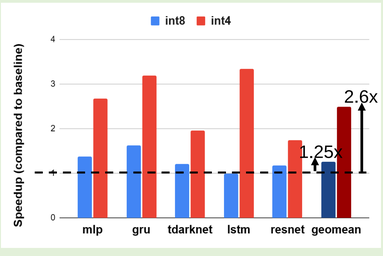
- Exploit dual-ported'ness of BRAMs to obtain 2 operands in 1 cycle
- Novel configurable processing element
- Two ways to program: FSM-based and Stored program
- Support One Operand Outside RAM (OOOR) operations

Pros and Cons

Increase compute throughput of the FPGA
Reduce routing requirement and data movement (hence energy)
Increase effective bandwidth (not limited by routing interface)
Precision agnostic. Great for DL!
Need to update tools, provide libraries, etc.
Bit-serial computation means higher latency

Results





Conclusion

FPGAs are commonly used for hardware acceleration of Deep Learning workloads. We propose DL-optimized FPGAs to reduce the performance gap between FPGAs and ASICs. The key idea is to integrate new hardware blocks to the FPGA that provide DL-specialized functionality, while still integrating seamlessly into traditional FPGA flow. With the abundance of DL applications, making DL-Optimized FPGAs is an attractive proposition.

References

1. Murray et al., VTR 8: High-Performance CAD and Customizable FPGA Architecture Modelling, ACM TRETS 2020
2. Arora et al., Tensor Slices to the Rescue: Supercharging ML Acceleration on FPGAs, ACM FPGA 2021
3. Arora et al., CoMeFa: Compute-in-Memory Blocks for FPGAs, IEEE FCCM 2022 (Best Paper Award)
4. Arora et al., Koios: A DL Benchmark Suite for FPGA Architecture and CAD Research, IEEE FPL 2021