Sandia National Laboratories

Exceptional service in the national interest

ATHENA

ENABLING HIGH SPEED PERFORMANCE ESTIMATES FOR NOVEL HARDWARE DESIGN SPACE EXPLORATION

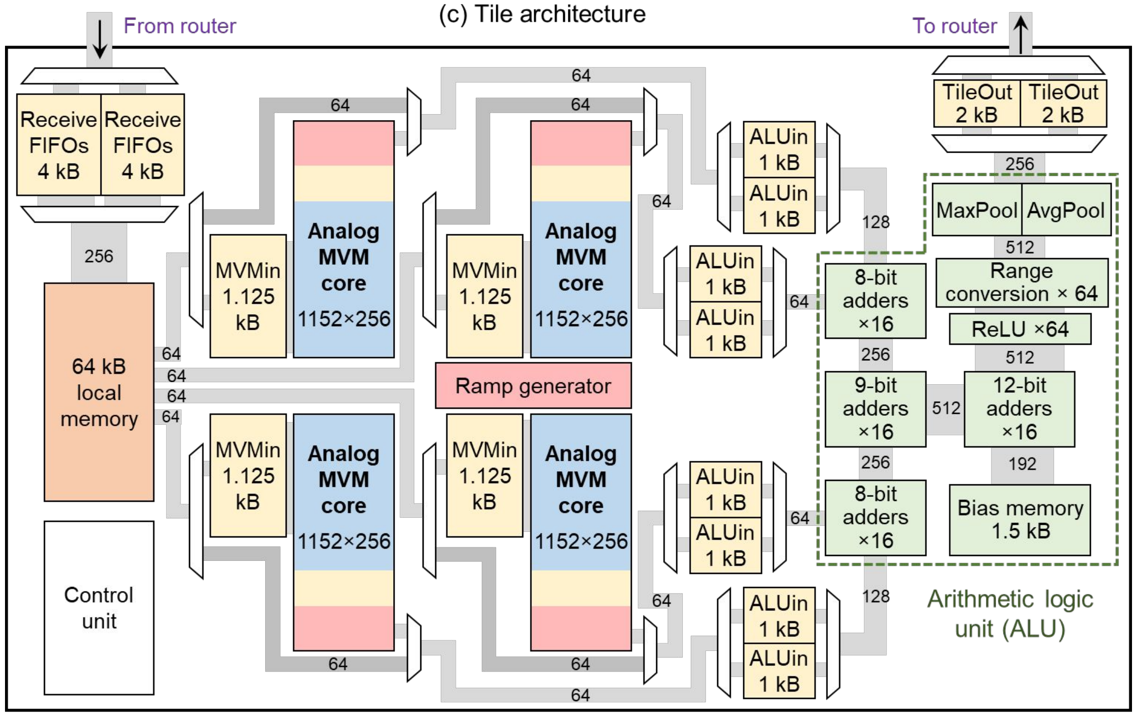
Mark Plagge, Suma Cardwell, and Clayton Hughes

SAND2021-12106 C

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

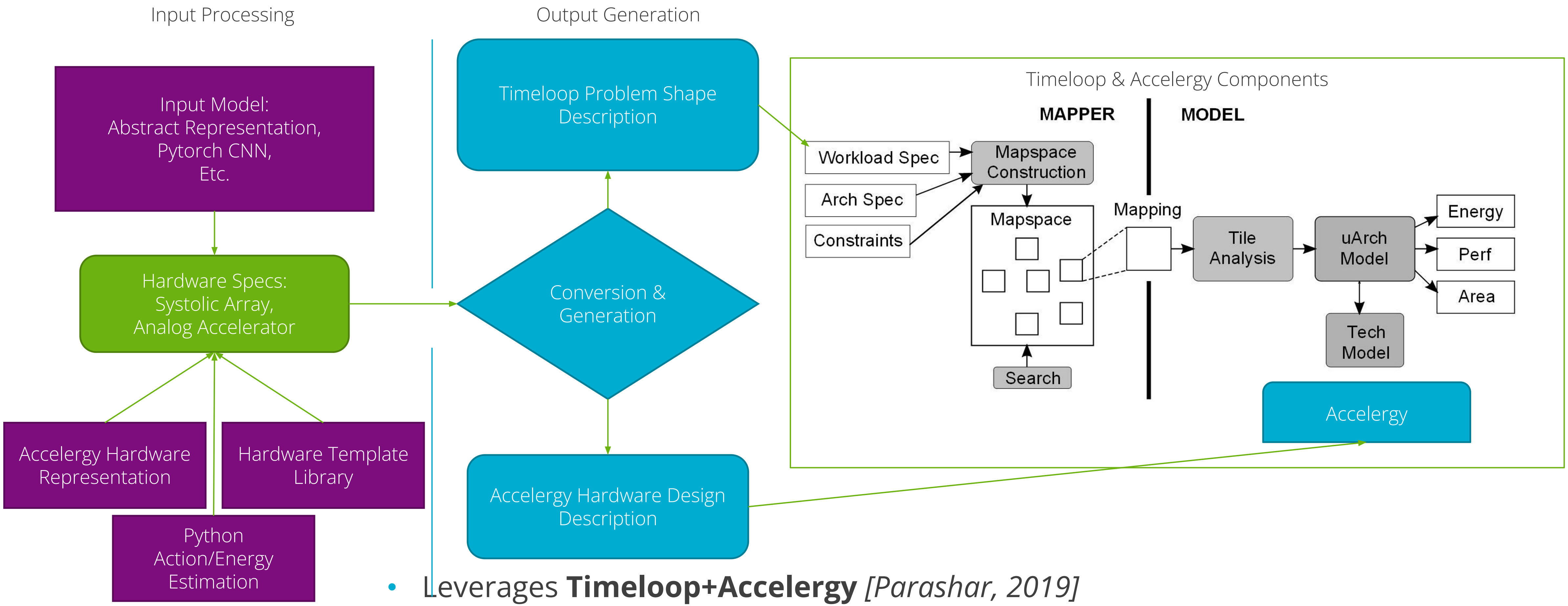
ATHENA – Rapid Performance Estimation for Novel Hardware

- Analog and neuromorphic accelerators have the potential to dramatically increase efficiency of many aspects of computing
- Analog devices are extremely low-energy when computing Matrix Vector Multiply operations
- There is a lack of fast and flexible benchmarking and design-space exploration tools for analog devices
- Yet there are many such tools for digital compute devices
- ATHENA: Leveraging analytical techniques to provide hardware performance estimates
- Currently supports the SONOS tiled MVM hardware architecture



SONOS Analog Accelerator Hardware Tile Architecture Design [Xiao, 2021]

ATHENA Plug-In



Input Processing

Output Generation

Input Model:
Abstract Representation,
Pytorch CNN,
Etc.

Hardware Specs:
Systolic Array,
Analog Accelerator

Accelergy Hardware Representation

Hardware Template Library

Python Action/Energy Estimation

Timeloop Problem Shape Description

Conversion & Generation

Accelergy Hardware Design Description

Timeloop & Accelergy Components

MAPPER

MODEL

Workload Spec

Arch Spec

Constraints

Mapspace Construction

Mapspace

Search

Mapping

Tile Analysis

uArch Model

Tech Model

Energy

Perf

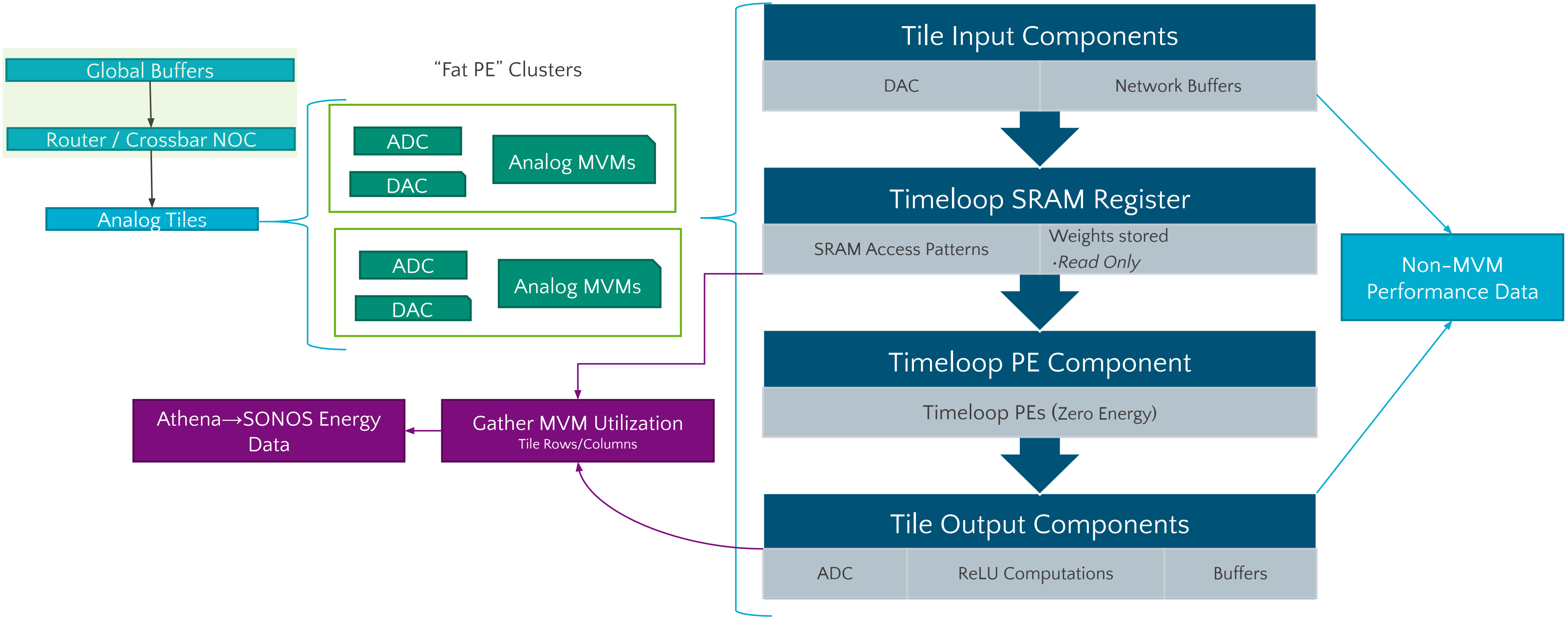
Area

Accelergy

- Leverages **Timeloop+Accelergy** *[Parashar, 2019]*
- ATHENA takes a problem layer and hardware description then:
  - Generates Accelergy energy table using Python
  - Generates Timeloop problem space
  - Runs Timeloop with hardware plugins
  - Collects and presents results

Analog Tiles as Dataflow Hardware

- ATHENA wraps the complex logic of an Analog cluster into a group of PEs and memory components
- Each “**Fat PE**” cluster contains dummy memory which is mapped to the analog array’s energy
  - Analog devices have energy costs based on the size of the compute
  - Timeloop only supports a fixed per-MAC energy cost
- To Timeloop the hardware appears as a set of PEs with zero energy cost behind a memory buffer



ATHENA Accuracy Compared to SONOS Simulator

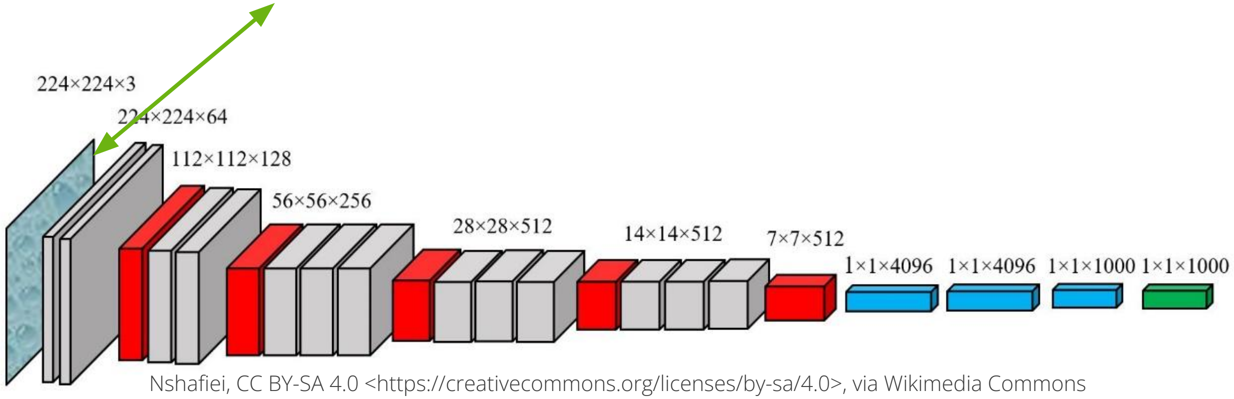
VGG 16

ATHENA MVM Compute Energy Accuracy

Layer	Total MACs	Athena	SONOS	Difference
Conv. 1	86 704 128	2.1431 pJ	1.0652 pJ	67.1968 %
Conv. 2	1 849 688 064	8.5201 pJ	3.7520 pJ	77.7058 %
Conv. 3	924 844 032	2.1295 pJ	2.1042 pJ	1.1940 %
Conv. 4	1 849 688 064	4.0181 pJ	3.9704 pJ	1.1940 %
Conv. 5	924 844 032	1.0647 pJ	1.0395 pJ	2.3951 %
Conv. 6	1 849 688 064	2.1295 pJ	2.0791 pJ	2.3951 %
Conv. 7	1 849 688 064	2.1295 pJ	2.0791 pJ	2.3951 %
Conv. 8	924 844 032	1.0647 pJ	1.0146 pJ	4.8186 %
Conv. 9	1 849 688 064	2.1295 pJ	2.0293 pJ	4.8186 %
Conv. 10	1 849 688 064	2.1295 pJ	2.0293 pJ	4.8186 %
Conv. 11	462 422 016	0.532 37 pJ	0.482 88 pJ	9.7503 %
Conv. 12	462 422 016	0.532 37 pJ	0.482 88 pJ	9.7503 %
Conv. 13	462 422 016	0.532 37 pJ	0.482 88 pJ	9.7503 %

ATHENA Tile Compute Energy Accuracy

Athena Tile	SONOS Tile Result	Number of Computations
22.196 pJ	21.548 pJ	86,704,128



Nshafiei, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

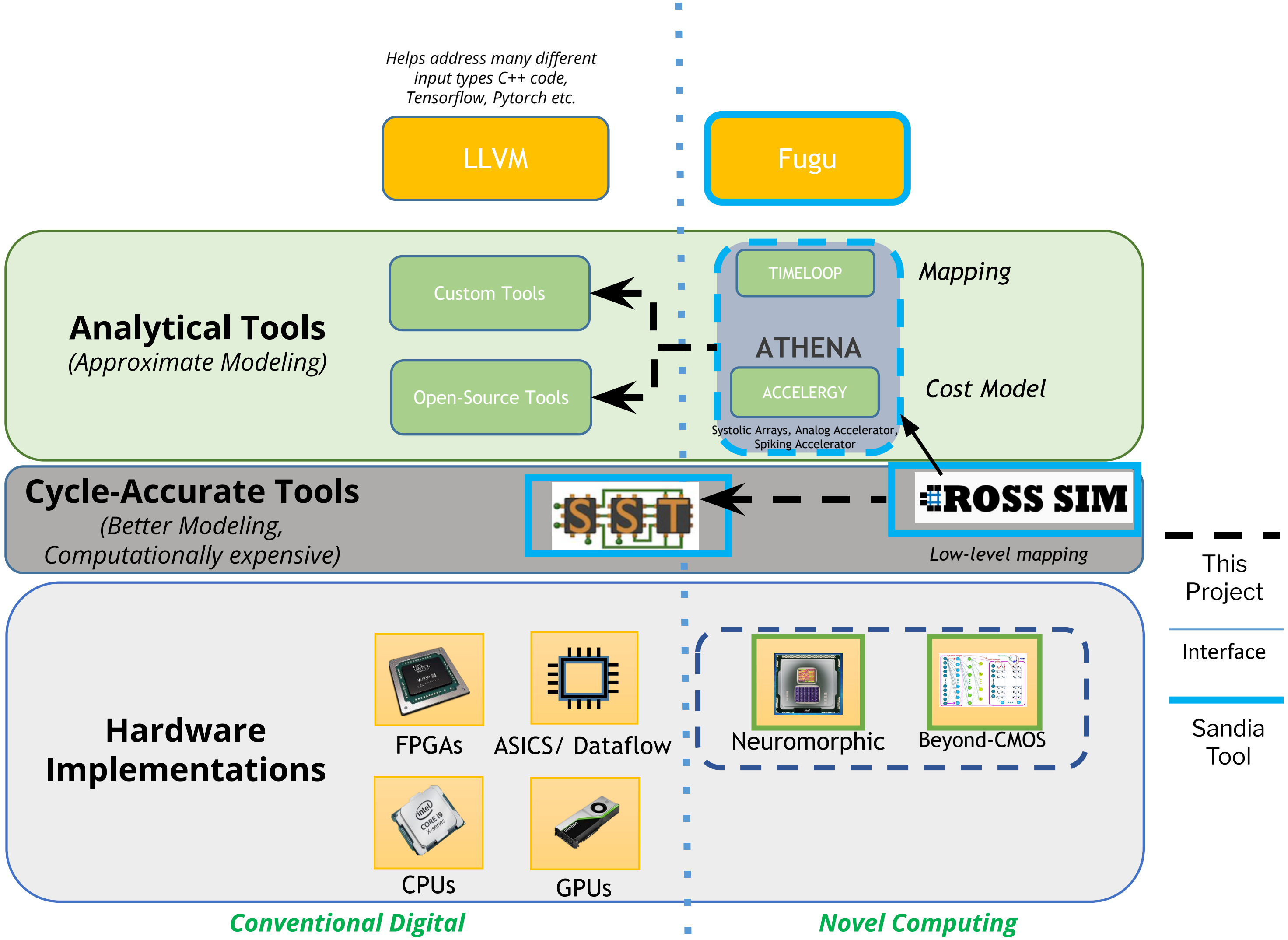
- Results are promising
- Accuracy of the total tile shows good potential
- SONOS simulator incorporates data from experimental devices by Infineon Tech.
- Comparing MVM compute energy shows a greater inaccuracy
  - This is attributed to ATHENA's naive implementation of mapping; The SONOS Simulator uses hand-mapped dataflows for improved performance

Future Work

- Support digital spiking neuromorphic hardware
  - *Model spiking activity*

- Support for other emerging devices

- Based on these preliminary results, we believe that ATHENA will be useful as part of a design-space-exploration tool for novel acceleration hardware:
  - Use ATHENA to search for efficient hardware designs and dataflow mapping
  - Leverage a highly detailed simulation model to gather more detailed results



Helps address many different input types C++ code, Tensorflow, Pytorch etc.
LLVM
Fugu
Analytical Tools
(Approximate Modeling)
Custom Tools
Open-Source Tools
TIMELOOP
ATHENA
ACCELERGY
Mapping
Cost Model
Systolic Arrays, Analog Accelerator, Spiking Accelerator
Cycle-Accurate Tools
(Better Modeling, Computationally expensive)
SST
CROSS SIM
Low-level mapping
Hardware Implementations
FPGAs
ASICS/ Dataflow
CPUs
GPUs
Neuromorphic
Beyond-CMOS
Conventional Digital
Novel Computing
This Project
Interface
Sandia Tool