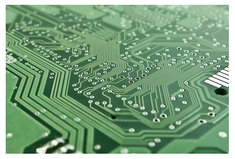
Sparseloop: An Analytical Approach To Sparse Tensor Accelerator Modeling

Yannan Nellie Wu, Po-An Tsai, Angshuman Parashar, Vivienne Sze, Joel S. Emer

1. Background & Motivation

Statistics Applications



Circuit Simulations

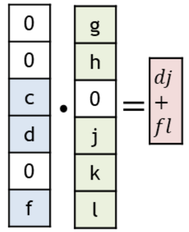
Networks

- **Sparse tensor algebra** is popular in many applications.
- Various sparse tensor accelerators are proposed to achieve higher speed and energy efficiency.
- **Important to <u>understand</u> and <u>rapidly explore</u> the diverse design space.**

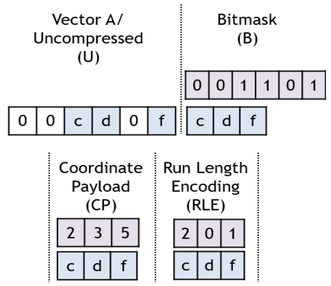
2. Design Space Classification

We classify common sparsity-aware acceleration techniques into three high-level <u>sparse acceleration features (SAFs)</u>

Example Workload



1. Representation format



2. Gating

processing steps ⟶

	1	*2*	*3*	*4*	*5*	*6*
Gating on B accesses based on A	read(0) ~~read(g)~~ ~~compute(0,g)~~	read(0) ~~read(h)~~ ~~compute(0,h)~~	read(c) read(0) compute(c,0)	read(d) read(j) compute(d,j)	read(0) ~~read(k)~~ ~~compute(0,k)~~	read(f) read(l) compute(f,l)

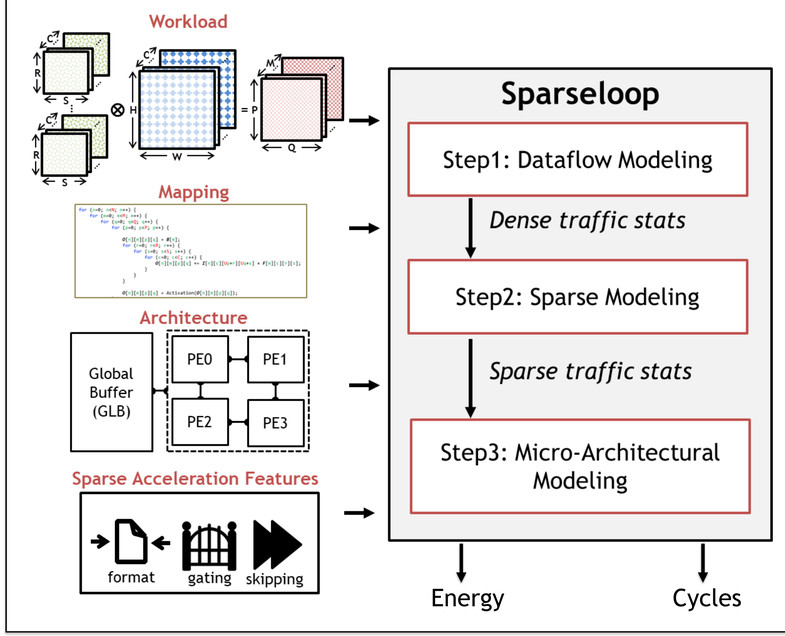
3. Skipping

processing steps ⟶

	1	*2*	*3*
Skipping on B accesses based on A	read(c) read(0) compute(c,0)	read(d) read(j) compute(d,j)	read(f) read(l) compute(f,l)

We need a <u>fast, accurate, and flexible</u> modeling framework to rapidly explore the various implementation choices

3. Modeling Framework



Modularized Modeling Process with Tractable Complexity

Module Actions

Step1: Derives the uncompressed data movement and the number of dense computes.

Step2: Analyzes the impact of each SAF based on statistical characterization of workload data (e.g., uniform distribution).

Step3: Calculates the final energy and cycle count based on microarchitectural details (e.g., technology node).

Tutorial Website: http://accelergy.mit.edu/sparse_tutorial.html

4. Experimental Results

- >2000x faster than cycle-level simulations.
- Accurately models well-known sparse tensor accelerators with 0.1% to 8% average error
- Provides the flexibility to evaluate and explore accelerators with various architecture topologies, dataflows, and SAFs, running workloads with various sparsity characteristics.

Example sweep on SAF implementations and workloads

Different energy breakdown across components for different workloads