Pipirima: Predicting Patterns in Sparsity to Accelerate Matrix Algebra

Ubaid Bakhtiar, Donghyeon Joo, and Bahar Asgari
University of Maryland, College Park

INTRODUCTION

Prevalence of sparsity across domains

Sparse matrix compression

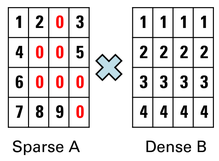
1	2	0	3
4	0	0	5
6	0	0	0
7	8	9	0

Compression → CSR Format

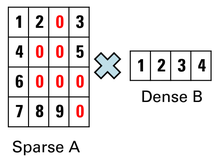
Key Benefits:
- ✓ Reduce storage overhead
- ✓ Improve performance
- ✓ Mitigate data movement

Sparse kernels

Sparse matrix-dense matrix multiplication **(SpMM)**

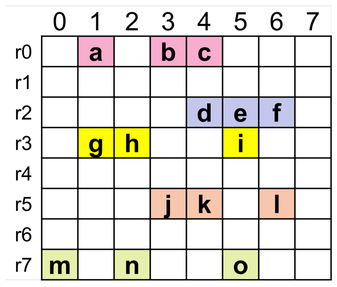


Sparse matrix-dense vector multiplication **(SpMV)**

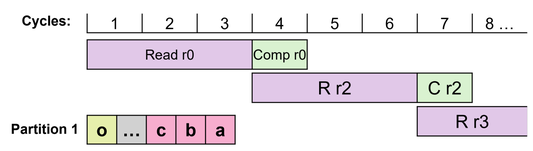


Key Challenges:
- Load imbalance
- Irregular memory accesses
- Decompression Overhead

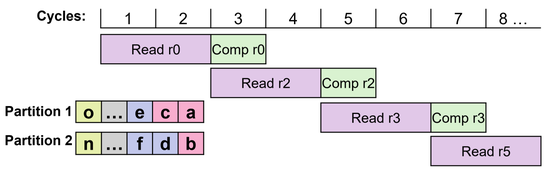
CHALLENGE: FINE-GRAINED LOAD IMBALANCE



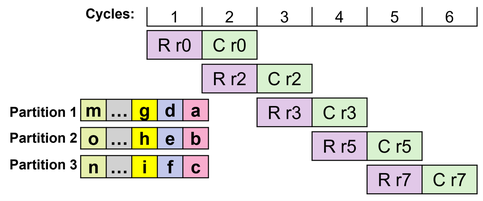
a. No distribution-Serial Placement of non-zero values results in bottlenecks



b. Sub-par parallelism due to non-zero distribution among two memory partitions

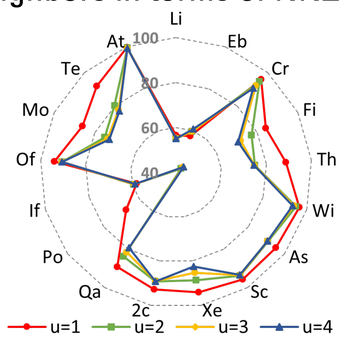


c. Maximum parallelism due to perfect non-zero distribution



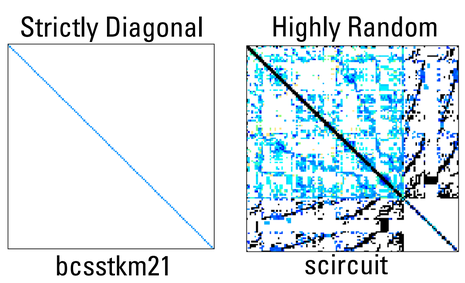
OBSERVATION: STRUCTURAL COHERENCE

Observation 1. On average **78-85%** of the rows are homogeneous to their last u neighbors in terms of NNZ.



Row homogeneity % for the last "u" neighbors of SuiteSparse workloads

Observation 2. The non-zeros location in strictly diagonal matrix is pre-determined and we can skip decompressing them.



Some sparse matrices exhibit strictly diagonal pattern

KEY INSIGHT

- ✓ To enable fine-grained load imbalance, Pipirima predicts the NNZ values for the current block based on history of previous blocks (Similar to branch predictors in CPUs).
- ✓ Using such a simple predictor, by just reading a single counter, Pipirima distribute them evenly as values and indices are streamed from memory.

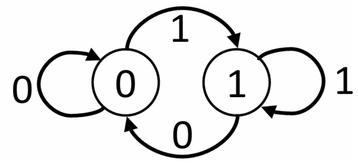
PREDICTION MECHANISM

D/R Matrix Predictor

It predicts if the sparse matrix has a random sparsity pattern or diagonal sparsity pattern.

State Diagram:

1: Random Matrix
0: Diagonal Matrix

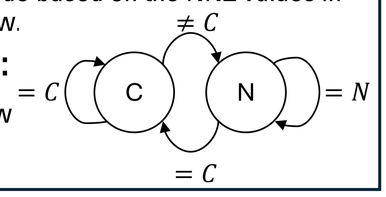


NNZ/row Predictor

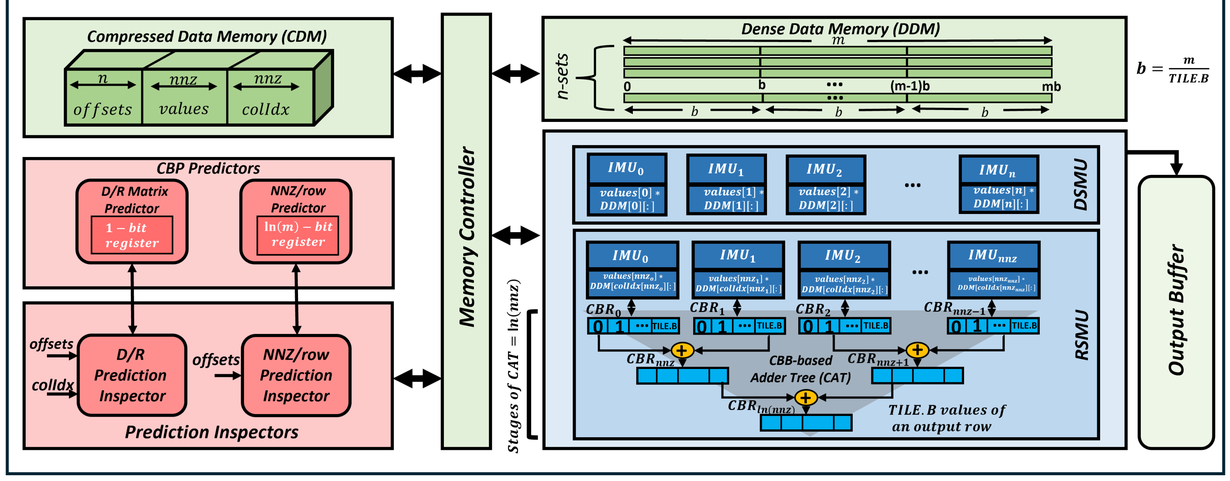
It predicts the number of non-zeros values per row. The decision is made based on the NNZ values in the subsequent row.

State Diagram:

C: current NNZ/row
N: new NNZ/row

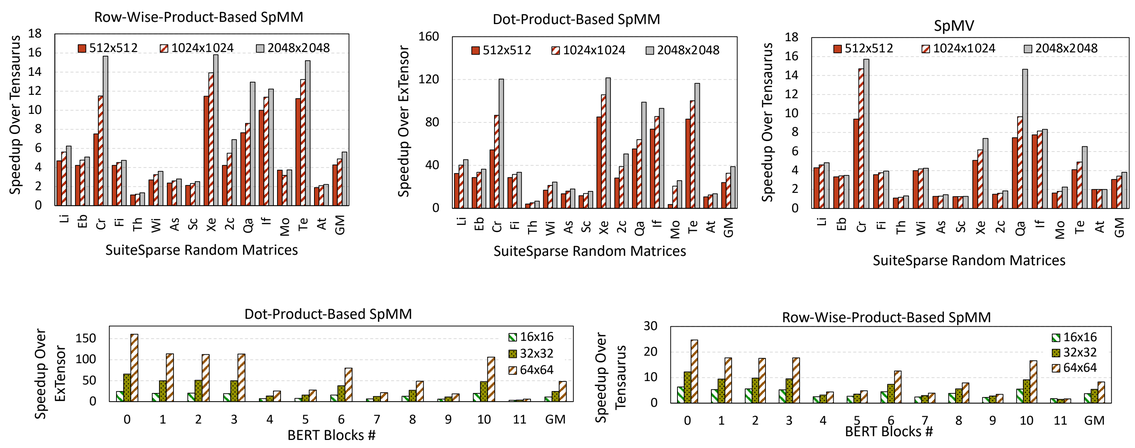


ARCHITECTURE OF PIPIRIMA



EVALUATION

Speedup



Mispredictions v. Density

Overhead of Prediction Components

Area and Power → Pipirima takes $5.621mm^2$ area and $544.9mW$ power with prediction related components taking only 0.15% and 0.54%

The 62nd Design Automation Conference (DAC)
San Francisco, CA, June 2025