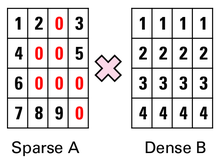
Chasoň: Supporting Cross HBM Channel Data Migration to Enable Efficient Sparse Algebraic Acceleration

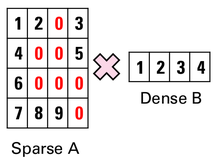
Ubaid Bakhtiar, Amirmahdi Namjoo, and Bahar Asgari
University of Maryland, College Park
Email: {ubaidb, namjoo, bahar} @ umd.edu

Sparse Algebra

SpMM kernel

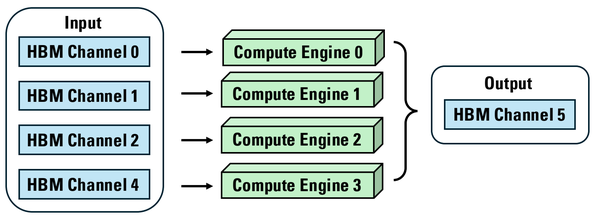


SpMV kernel



Sparse Algebraic Streaming Accelerators

- Data-driven execution model of streaming accelerators
- Maximize data transfer and computational throughput
- Compute engines' data scheduling in the HBM channels

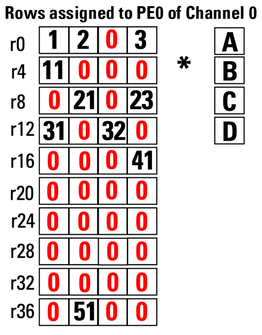


Prevalence of sparsity across domains

Challenges

PE-aware Non-zero Scheduling (state-of-the-art)

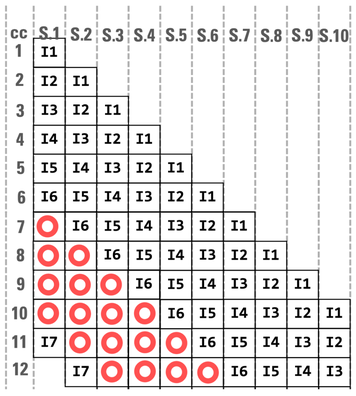
Rows assigned to PE0 of Channel 0



PE0 of Channel 0 Instructions:

```
I1: r0_op1  = 1A
I2: r4_op1  = 11A
I3: r8_op1  = 21B
I4: r12_op1 = 31A
I5: r16_op1 = 41D
I6: r36_op1 = 51B
I7: r0_op2  = 2B   +r0_op1
```

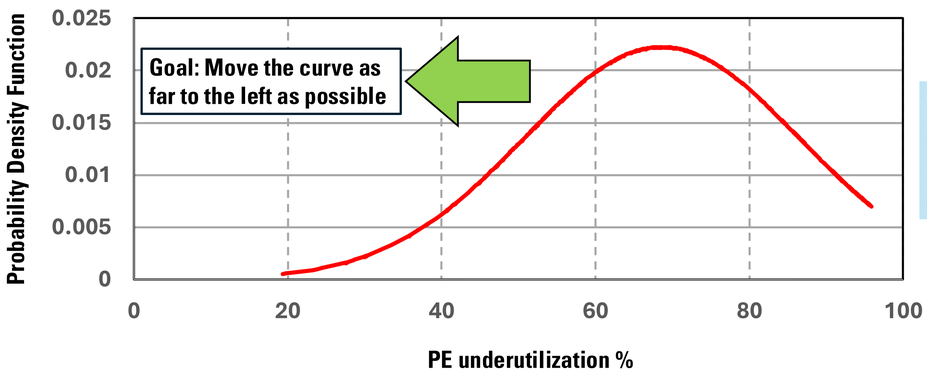
RAW



Legend:
cc: cycle S.x: Stage x
NOP

Throughput: **0.60** non-zeros per cycle

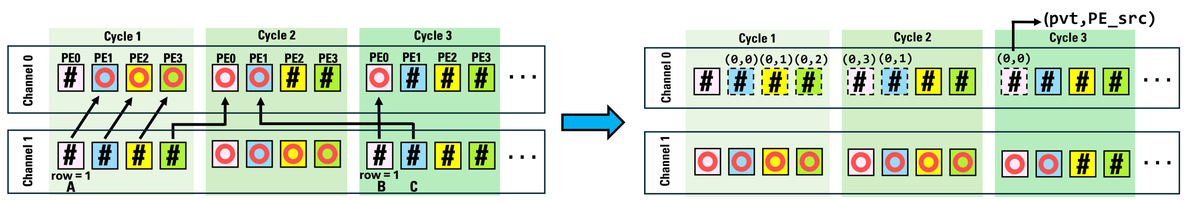
PE0 underutilization: $\frac{4\ stalls}{10}$ = **40%**

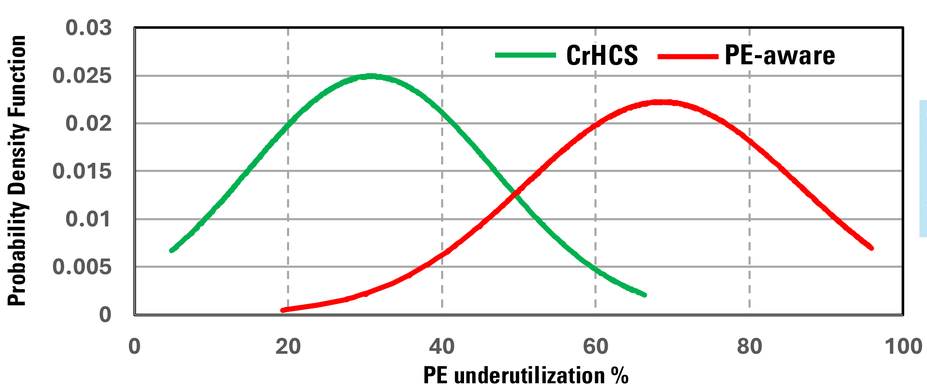


Caption. PE-aware non-zero scheduling (state-of-the-art) results in high PE underutilization % on 800 SuiteSparse as shown in this figure (lower is better).

Our Proposed Scheduling - Cross-HBM Channel OoO Scheduling (CrHCS)

Mechanism

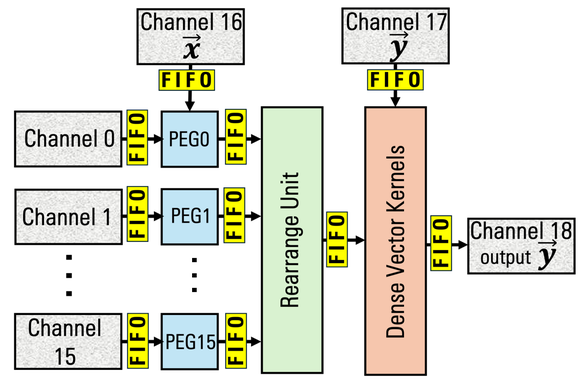




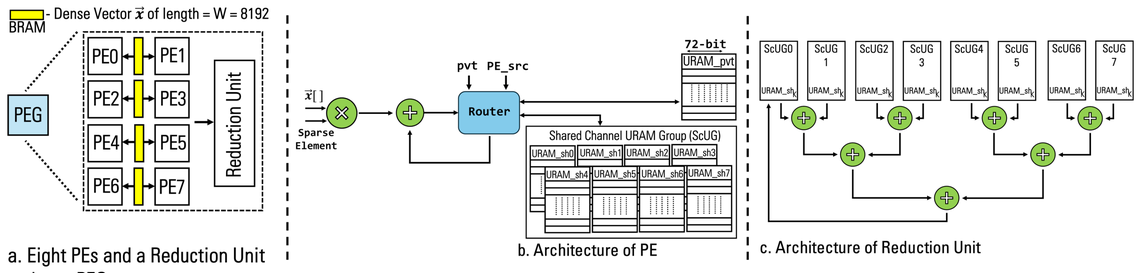
Caption. PE underutilization % in 800 SuiteSparse matrices of CrHCS and PE-aware (lower is better). CrHCS is an efficient scheduling scheme over PE-aware non-zero scheduling.

Our Proposed Architecture - Chasoň

High-level Overview



Architecture of PEG



a. Eight PEs and a Reduction Unit makes a PEG

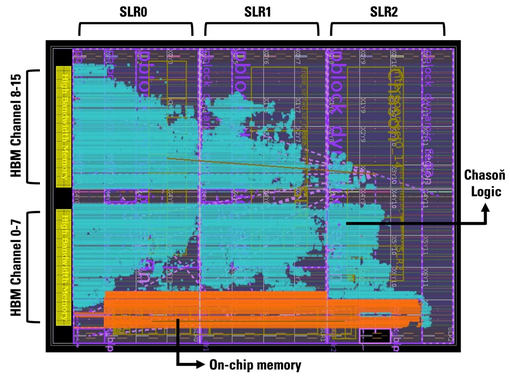
b. Architecture of PE

c. Architecture of Reduction Unit

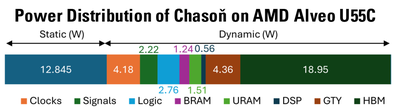
Evaluation

Hardware Implementation

Layout of Chasoň on AMD Alveo U55C



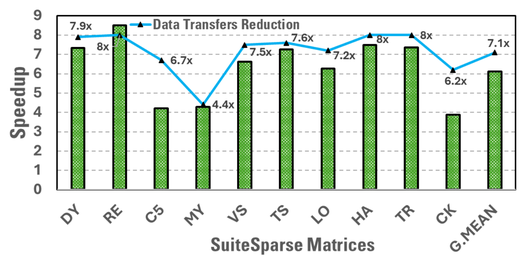
Power Distribution of Chasoň on AMD Alveo U55C

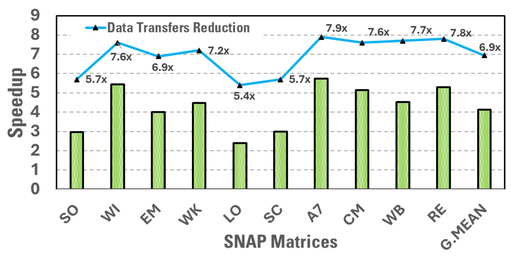


AMD Alveo U55C Resource Consumption for Chasoň & Serpens

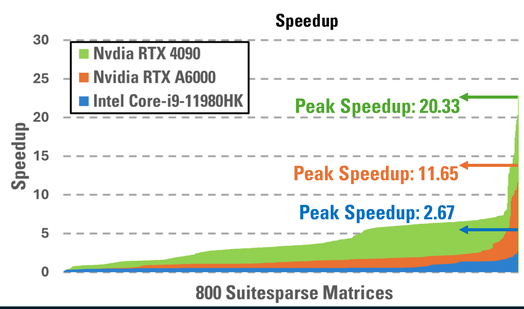
	Architecture	
	Serpens	Chasoň
LUT	219K (16%)	346K (26%)
FF	252K (9.6%)	418K (16%)
DSP	798 (9.6%)	1254 (13%)
BRAM18K	1024 (28%)	1024 (28%)
URAM	384 (40%)	512 (52%)

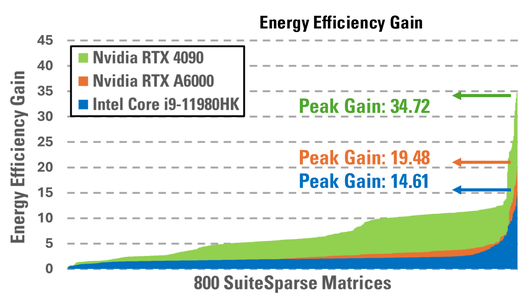
Speedup over Serpens (state-of-the-art SpMV accelerator)





Speedup over GPU (cuSPARSE) and CPU (Intel Math Kernel Library)





Paper

Artifact

Slides

Supported by:
DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA
NSF

58th IEEE/ACM International Symposium on Microarchitecture (MICRO 2025)
Seoul, South Korea