ExTensor: An Accelerator for Sparse Tensor Algebra

Kartik Hegde[1], Hadi Asghari-moghaddam[1], Michael Pellauer[2], Neal Crago[2], Aamer Jaleel[2], Edgar Solomonik[1], Joel Emer[2,3] ,Christopher W. Fletcher[1]

[1]University of Illinois at Urbana-Champaign
[2]NVIDIA
[3]MIT

1. DENSE/SPARSE TENSOR ALGEBRA

Rapidly growing in importance across Computer Science!
From Deep Learning to Recommendation Systems.

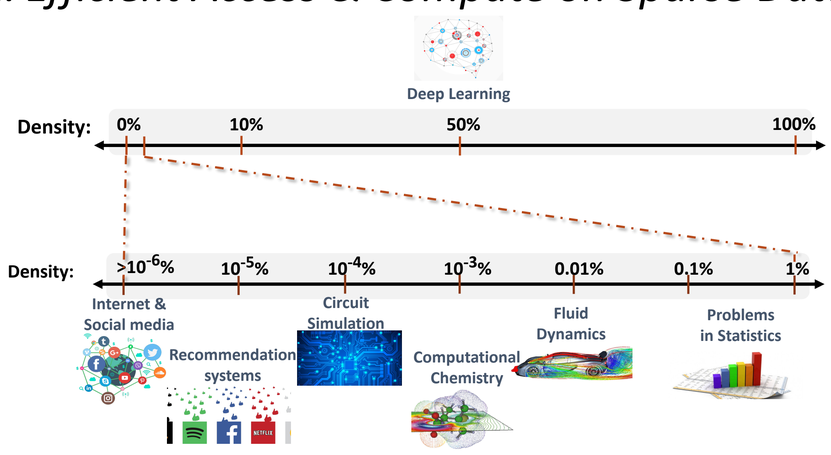
Core Challenges

1. *Diversity of Tensor Algebra Kernels*.

$$Z_i = \sum_i A_i B_i$$
$$Z_{ij} = A_{ij} \sum_k C_{ik} D_{kj} \qquad Z_{ijk} = \sum_{jl} A_{ijl} B_{kl}$$
$$Z_{ii} = \sum A_{ik} B_{ki} \qquad Z_{ij} = \sum_k A_{ijk} B_k$$
$$O_{xy} = \sum_{rs} I_{(x+r)(y+s)} F_{rs}$$

Significantly different compute characteristics!

2. *Efficient Access & Compute on Sparse Data Structures*



Need to Store them compressed!

Extensor

1. Fully ***programmable*** for Generalized Tensor Algebra.
2. ***Hierarchical*** Intersection to eliminate ineffectual work.
3. Highly ***Optimized Intersection*** Algorithm.

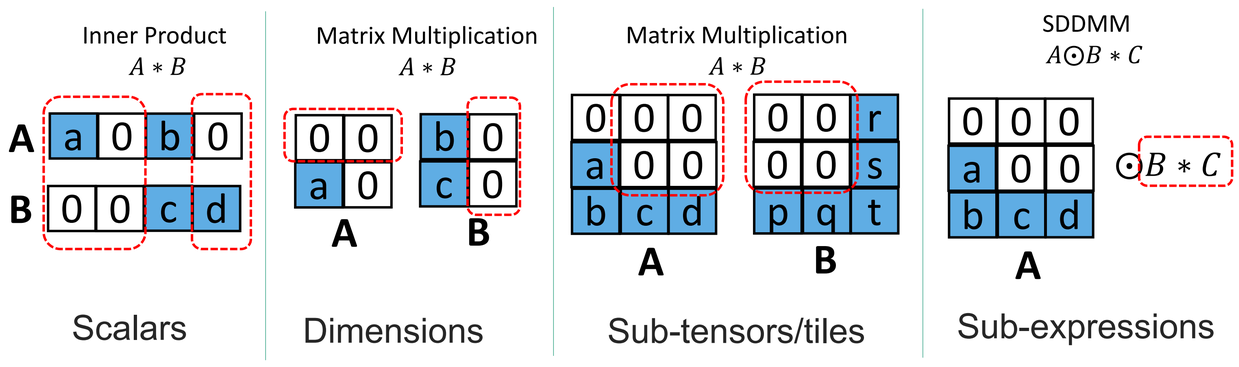
2. HIERARCHICAL INTERSECTION

Intersection: Identifying regions of overlapped non-zero regions while multiplying two sparse data-structures.

Basic Opportunity: $x * 0 = 0$

Key Insight

Above opportunity is applicable at different abstractions!



Increasing Intersection Granularity

Extensor

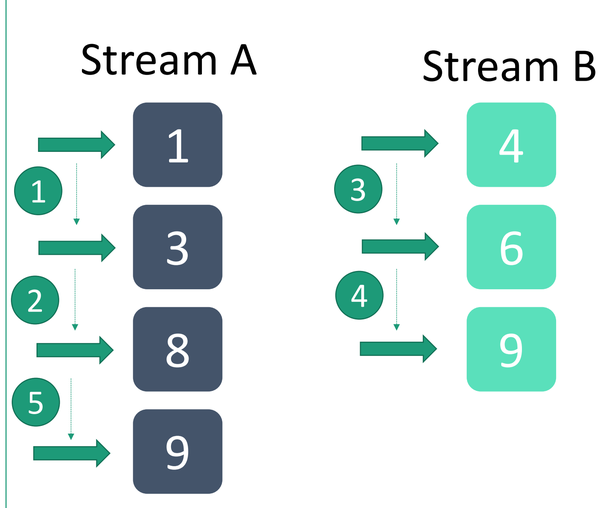
1. Combine the above opportunities **hierarchically**!
2. Leads to extremely effective ineffectual work skipping at **different granularities**.
3. Leverage the **metadata** of compressed representations to detect opportunities.

3. OPTIMIZED INTERSECTION METHODS

Motivation

Fast intersection is key to achieve high performance in Sparse Tensor Algebra.

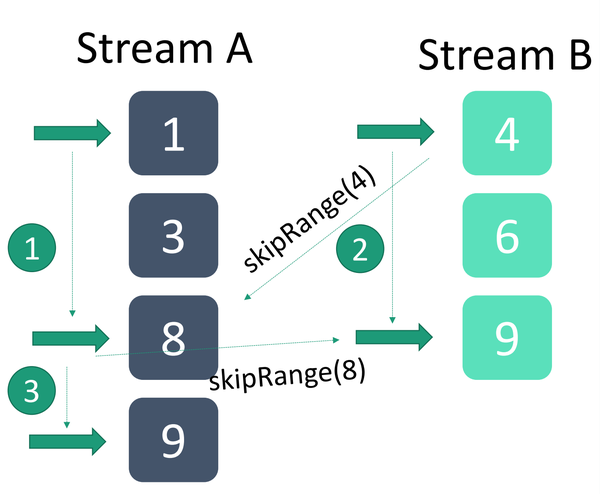
Naive Intersection



Walk step-by-step comparing the co-ordinates

O(|StreamA ∪ StreamB|)

ExTensor Intersection



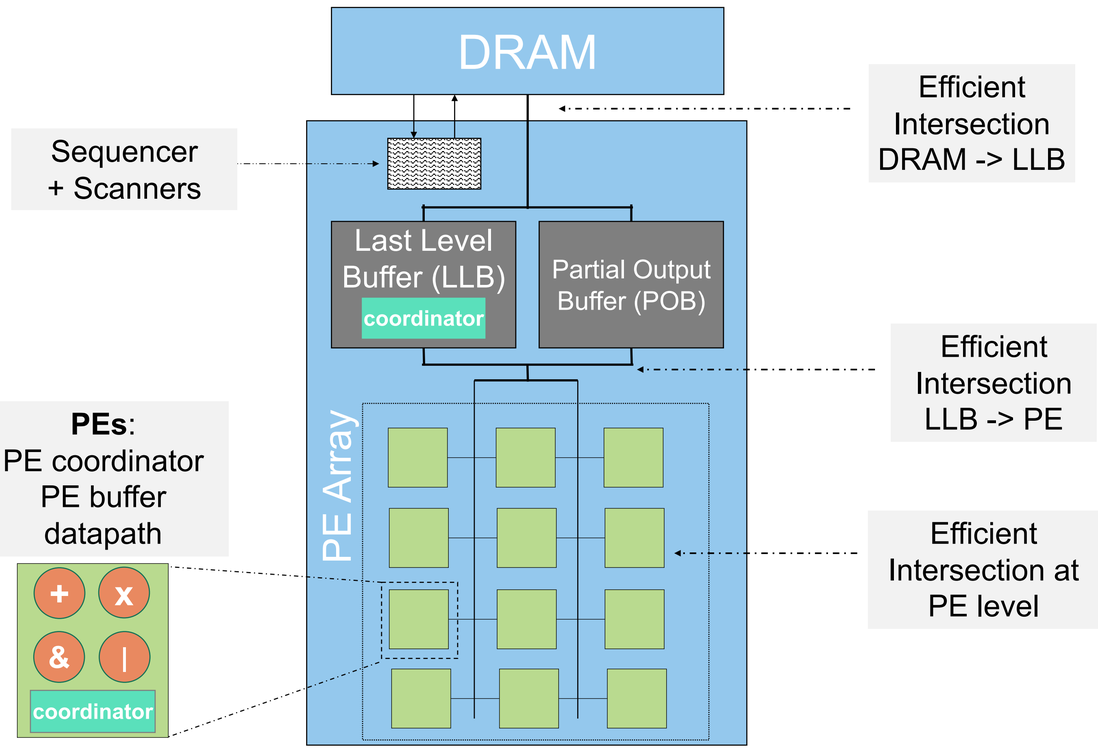
Each Stream uses skipRange() to skip a range.

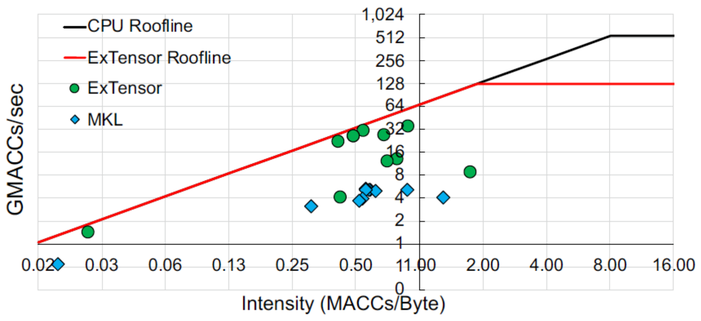
O(|StreamA ∩ StreamB|)

Extensor

1. Specialized hardware to implement optimized intersection.
2. Hierarchically replicate the intersection logic.

4. EXTENSOR ACCELERATOR DESIGN





Superior Resource Utilization
Via efficient intersection

Kernel	Improvement Over CPU
SpMSpM	3.4x
SpMM	1.3x
TTV	2.8
TTM	24.9
SDDMM	2.7x

Evaluation