Fine-Grained Accelerators for Sparse Machine Learning Workloads

ASP-DAC 2017

Asit K. Mishra, Eriko Nurvitadhi, Ganesh Venkatesh, Jonathan Pearce, Debbie Marr

Accelerator Architecture Lab
Intel Corp.

Disclaimer: The views expressed in this talk are those of the speaker and not his employer.

Executive Summary

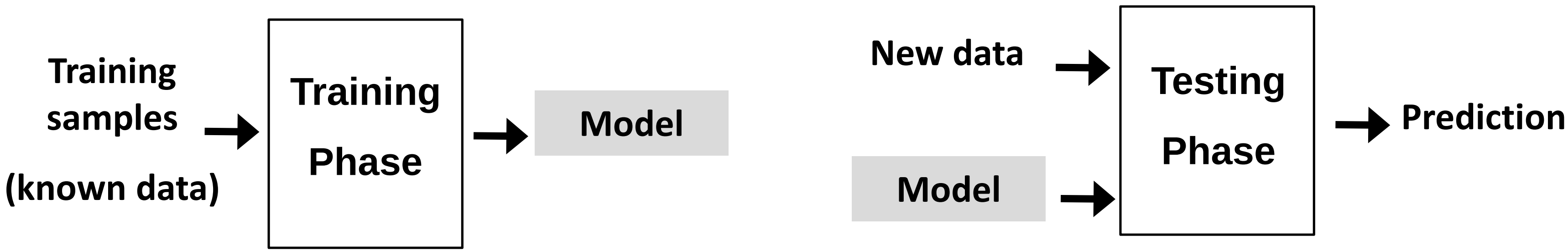
- Motivation
  - Data analytics growing in importance
  - They rely on machine learning (ML) algorithms
  - Working on datasets that are sparse (texts, ratings)

- This work: accelerate sparse ML workloads
  - Characterized ML workloads → low IPC, mem & branch mispredict stalls, high $ miss rate
  - Proposed HW accelerator → 4-13x speedups and 9-17x better energy over CPU, with small area

Talk Outline

- Executive summary
- <u>Sparse Machine Learning</u>
- Sparse matrix processing
- Characterization study
- Proposed hardware accelerator
- Related work + summary

Machine Learning for Data Analytics



Data represented as Matrix

Misplaced top-level domain (TLD)

URL	Reputation		Feature 0	Feature 1	...
www.ebay.*com*.phishy.biz	malicious	**sample0**	True		
www.ebay.com	normal	**sample1**	False		
		...	…	…	…

Sparse

$\underbrace{\hspace{6em}}_{\textbf{High-dimensional}}$

Many Real World Datasets Are Large, Sparse, and High-Dimensional

Examples from datasets we studied

Name	Avg. length	n=Max. length	l = # Samples	% Sparsity	Size on disk
E2006	1242	150K	16K	0.83%	485MB
RCV	74	42K	677K	0.15%	1.2GB
Webspam Unigram	86	255	245K	33.38%	268MB
Webspam Trigram	3.7K	399K	245K	0.04%	17GB
Gamevideo	221	1K	97K	22%	225MB
URL	117	2.6M	1.6M	0.003%	1.5GB
CriteoLabs	33	25.2M	32M	0.0001%	25GB
MovieLens	70K users,10K movies 10M ratings				253MB

Talk Outline

- **Executive summary**
- **Sparse Machine Learning**
- <u>**Sparse matrix processing**</u>
- **Characterization study**
- **Proposed hardware accelerator**
- **Related work + summary**

Matrix Formats: CSR vs. CSC

A

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

rowptr

row0 (3,2)
row1 (7,0)
row2 (4,2), (8,3)
row3 (6,0), (5,1), (3,2)
row4 (2,0), (1,3)
row5 (8,3)

colptr

col0 (7,1), (6,3), (2,4)
col1 (5,3)
col2 (3,0), (4,2), (3,3)
col3 (8,2), (1,4), (8,5)

A matrix example

Compressed Sparse Row (CSR)

Compressed Sparse Column (CSC)

Rows = samples

Columns = features

Good for operation on samples

Good for operation on features

Example Matrix Operations

spMdV_csr:

Row-oriented sparse matrix * dense vector

$$\overset{A}{\begin{bmatrix} 0 & 0 & 3 & 0 \\ 7 & 0 & 0 & 0 \\ 0 & 0 & 4 & 8 \\ 6 & 5 & 3 & 0 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 8 \end{bmatrix}} * \overset{x}{\begin{bmatrix} 1 \\ 2 \\ 4 \\ 3 \end{bmatrix}} = \overset{y}{\begin{bmatrix} 12 \\ \\ \\ \\ \\ \\ \end{bmatrix}}$$

...

$$\begin{bmatrix} 0 & 0 & 3 & 0 \\ 7 & 0 & 0 & 0 \\ 0 & 0 & 4 & 8 \\ 6 & 5 & 3 & 0 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 8 \end{bmatrix} * \begin{bmatrix} 1 \\ 2 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 12 \\ 7 \\ 40 \\ 28 \\ 5 \\ 24 \end{bmatrix}$$

Irregular reads on x

spMspV_csc:

Column-oriented sparse matrix * sparse vector

$$\overset{A}{\begin{bmatrix} 0 & 0 & 3 & 0 \\ 7 & 0 & 0 & 0 \\ 0 & 0 & 4 & 8 \\ 6 & 5 & 3 & 0 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 8 \end{bmatrix}} * \overset{x}{\begin{bmatrix} 0 \\ 0 \\ 4 \\ 8 \end{bmatrix}} = \overset{y}{\begin{bmatrix} 12 \\ 0 \\ 16 \\ 12 \\ 0 \\ 0 \end{bmatrix}}$$

$$\begin{bmatrix} 0 & 0 & 3 & 0 \\ 7 & 0 & 0 & 0 \\ 0 & 0 & 4 & 8 \\ 6 & 5 & 3 & 0 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 8 \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ 4 \\ 8 \end{bmatrix} = \begin{bmatrix} 12 \\ 0 \\ 80 \\ 12 \\ 8 \\ 64 \end{bmatrix}$$

Irregular reads and writes on y

spMdV_csc:

Scale matrix using scaling factors in x, then update y

$$\overset{A}{\begin{bmatrix} 0 & 0 & 3 & 0 \\ 7 & 0 & 0 & 0 \\ 0 & 0 & 4 & 8 \\ 6 & 5 & 3 & 0 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 8 \end{bmatrix}} * \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \end{bmatrix} \text{ x (scaling factors)}$$

$$= \begin{bmatrix} 0 & 0 & 3 & 0 \end{bmatrix} \text{ y}$$

...

$$\overset{A}{\begin{bmatrix} 0 & 0 & 3 & 0 \\ 7 & 0 & 0 & 0 \\ 0 & 0 & 4 & 8 \\ 6 & 5 & 3 & 0 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 8 \end{bmatrix}} * \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \end{bmatrix} \text{ x (scaling factors)}$$

$$= \begin{bmatrix} 24 & 5 & 12 & 25 \end{bmatrix} \text{ y}$$

Irregular reads and writes on y

Talk Outline

- **Executive summary**
- **Sparse Machine Learning**
- **Sparse matrix processing**
- **<u>Characterization study</u>**
- **Proposed hardware accelerator**
- **Related work + summary**

Methodology

- System Under Study
  - 2.7 GHz Intel Ivy Bridge Server (E5-2679 v2)
    - 24 cores, 32KB I-cache, 32KB D-cache,
    - 256 KB private L2 cache, 30 MB shared L3
    - 128 GB DDR3 memory, 60 GB/s max mem bandwidth
- Dataset
  - Real datasets, shown in earlier slide
- Tools
  - Vtune and gprof for hotspot characterizations
  - Sniper simulator to get cache statistics
  - McPat for energy modeling

Workloads and Identified Hotspots

Application	Type	Hot code	% Time
Sparse PCA	Dim. reduction	`SpVSpV`	99%
Kernelized SVM classification	Classification	`SpMSpV`	96%
Linear SVM classification	Classification	`SpMDV, SpVDV`	99%
Logistic regression	Classification	`SpMDV, SpVDV`	98%
Kernelized SVM regression	Regression	`SpMSpV`	94%
Linear SVM regression	Regression	`SpMDV, SpVDV`	99%
SLIM	Recom. engine	`SpMDV`	88%
ALS	Recom. engine	`SpMDV`	92%
K-means	Clustering	`SpVDV`	90%

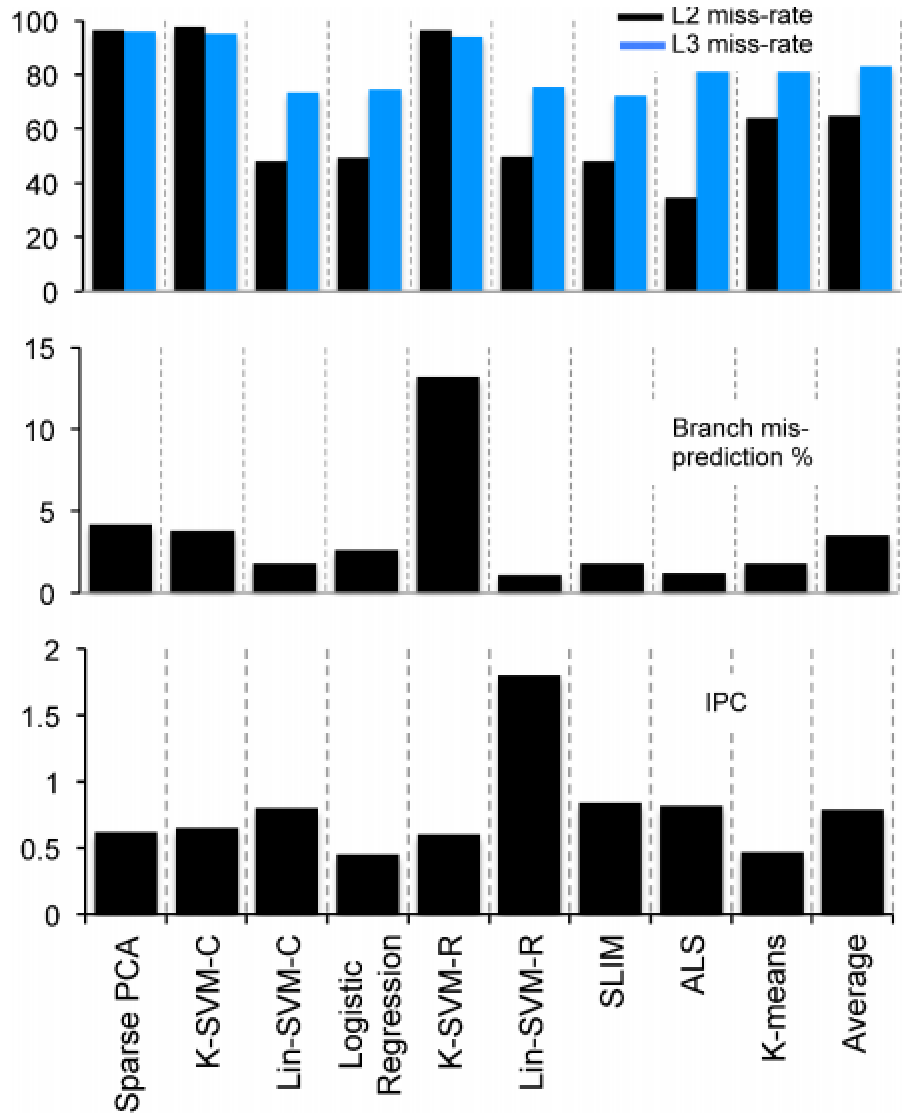
Majority of time spent on sparse matrix/vector ops

Application Characteristics

High cache miss rate

High Branch Misprediction

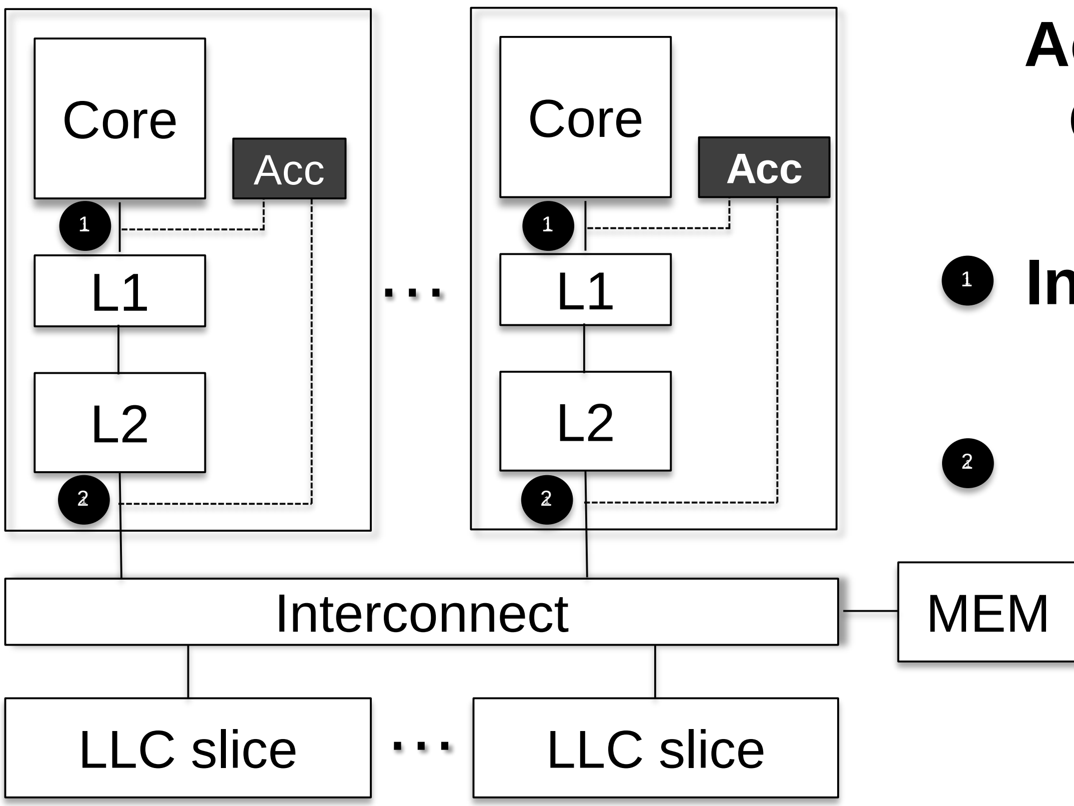
Low IPC



Talk Outline

- **Executive summary**
- **Sparse Machine Learning**
- **Sparse matrix processing**
- **Characterization study**
- **<u>Proposed hardware accelerator</u>**
- **Related work + summary**

System Architecture

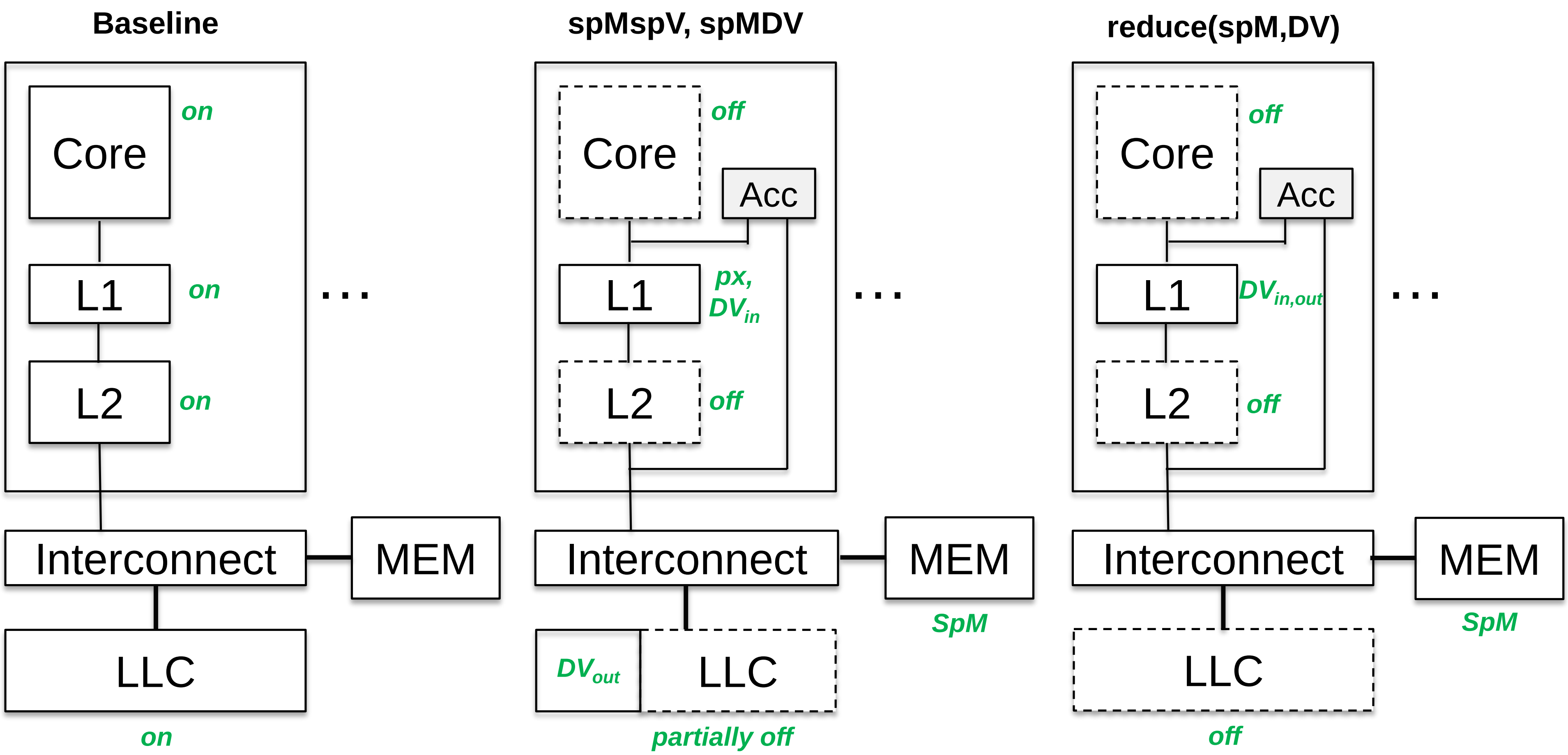


Accelerator Tightly Connected to CPU

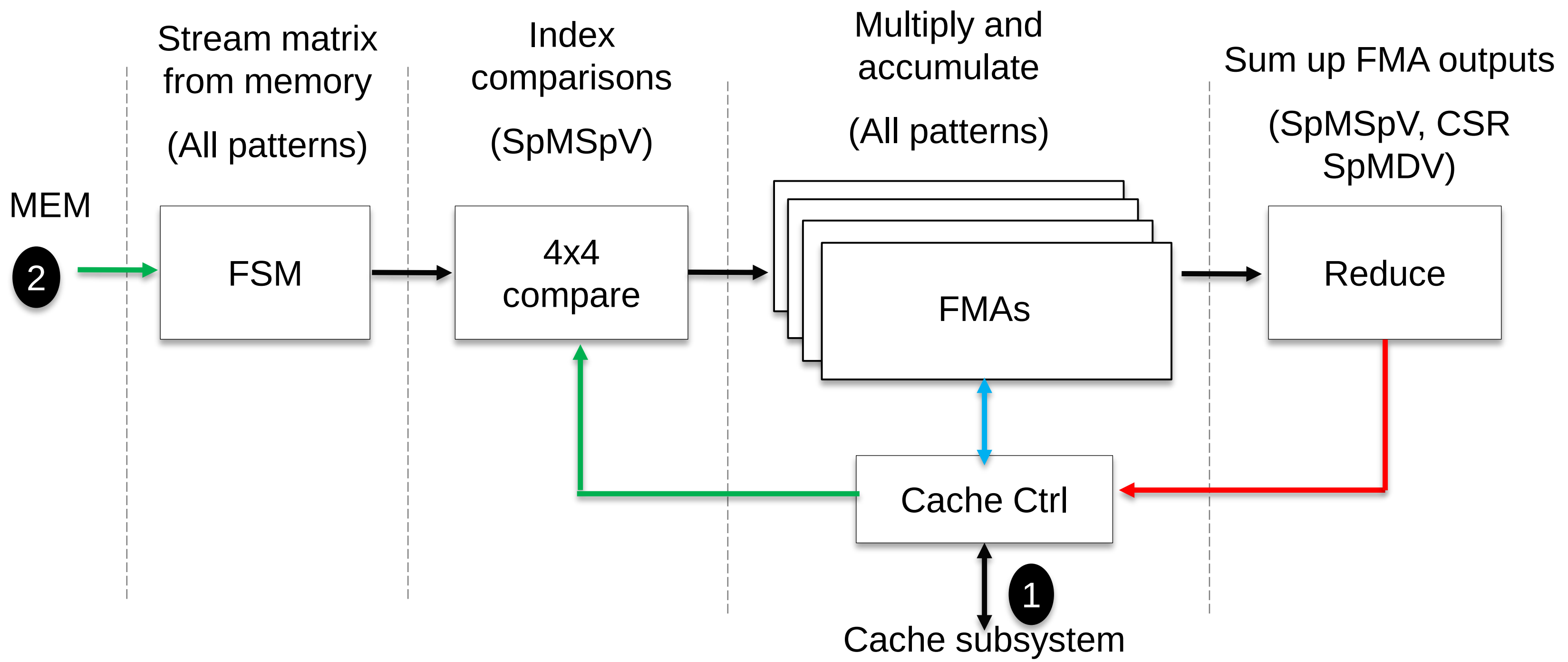
1 **Interface to L1 cache**

2 **Direct channel to external memory**

How to improve efficiency: custom config for each matrix ops



Accelerator Internal



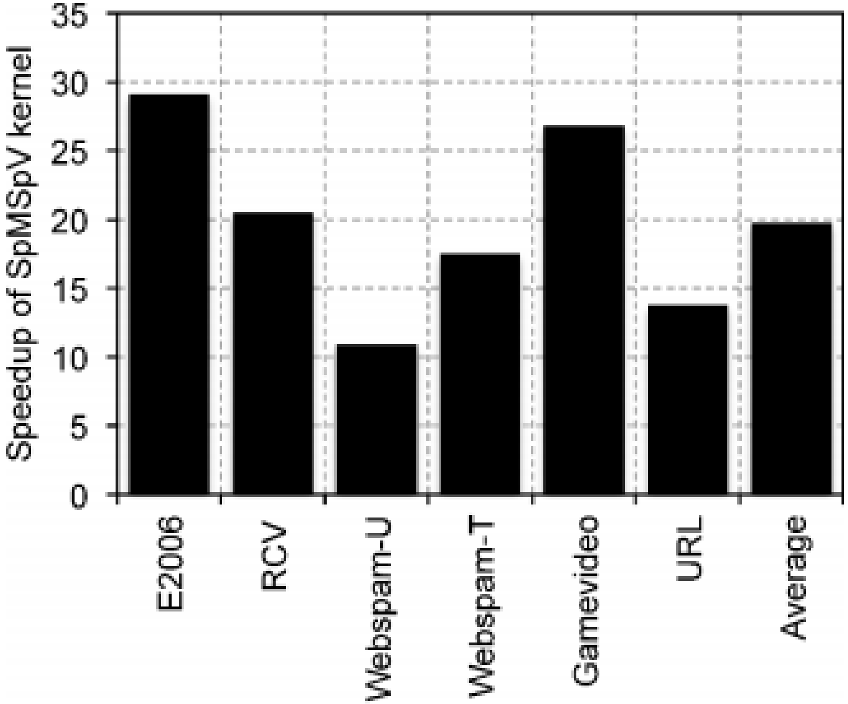
Support matrix operations used in ML workloads under study

(See details in the paper)

Results

~20x speedup for SpMSpv

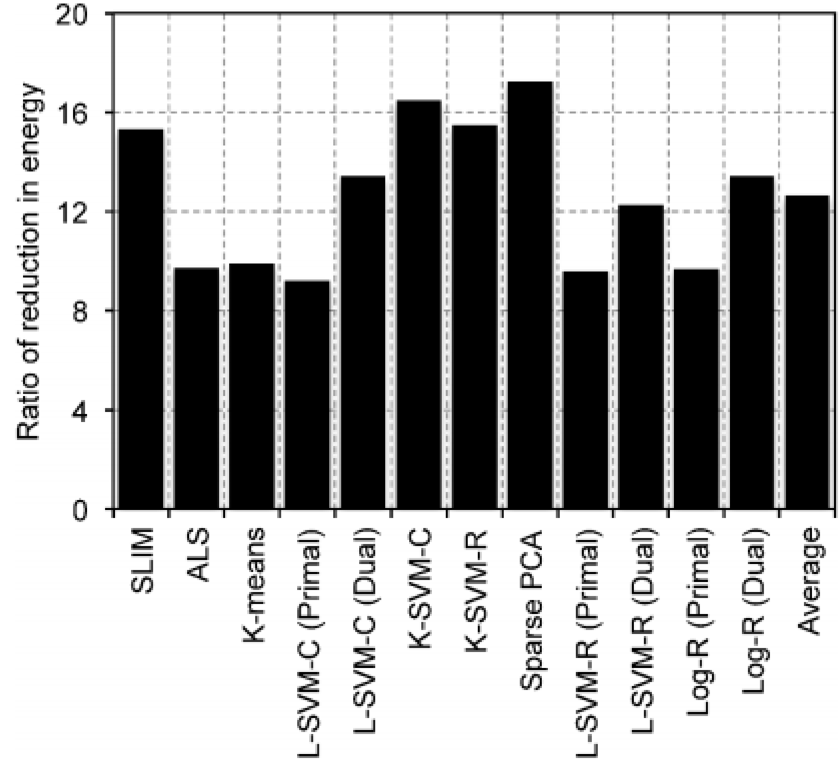
(K-SVM, sparse PCA)



Due to parallel index matching & custom datapath

SpMDV is bandwidth bound, perf limited by mem system

Energy efficiency improvements



~12x avg better energy

From turning off core and parts of mem subsystem

Talk Outline

- Executive summary
- Sparse Machine Learning
- Sparse matrix processing
- Characterization study
- Proposed hardware accelerator
- <u>Related work + summary</u>

Related Work

- Sparse matrix & sparse ML accelerator
  - Many proposals target only 1 sparse matrix format/op
  - Our previous work on sparse ML accelerator did not tightly integrate accelerator blocks with CPU
- Other ML accelerators
  - Many proposals for individual workloads
  - Many proposals for neural networks and/or dense data

Summary

- Sparse ML growing in importance
  - Sparsity from unstructured data (e.g., texts, ratings)

- We characterized various sparse ML workloads
  - Most runtime spent on sparse matrix op hotspots

- We proposed HW accelerator for these matrix ops
  - Tightly coupled with CPU and mem system
  - Improve efficiency dramatically (performance, energy)