Towards an End-to-End Processing-in-DRAM Acceleration of Spectral Library Search

Tianyun Zhang*, Eric Tang*, Farzana Ahmed Siddique†, Kevin Skadron†, Franz Franchetti*
*Electrical and Computer Engineering Department Carnegie Mellon University, Pittsburgh, PA, USA
†Computer Science Department University of Virginia, Charlottesville, VA, USA
{tianyun2, erictang, franzf}@andrew.cmu.edu, {farzana, skadron}@virginia.edu

***Abstract*—This work explores accelerating spectral library searches, a key mass spectrometry (MS) workload, using processing-in-memory (PIM) architectures through an end-to-end, co-designed approach. We apply signal processing and approximate computing techniques for pre-filtering MS data and implement a sum of absolute differences (SAD) algorithm optimized for PIM to compare spectral similarity. Our methodology is evaluated using a DRAM-based PIM simulator and compared against traditional CPU implementations. While initial results with small datasets favor CPUs, our analysis indicates potential benefits for PIM with larger, more realistic proteomics datasets. This work represents an initial step towards investigating PIM acceleration for MS applications.**

***Index Terms*—processing-in-memory, near-data processing, application-specific acceleration, bioinformatics, mass spectrometry, spectral library search**

I. Introduction

MS-based proteomics generates vast amounts of data, with modern instruments producing millions of spectra per experiment. Spectral library search is a crucial workload in protein identification. Traditional compute-centric architectures struggle with data movement between memory and processors, a challenge that grows with increasing dataset sizes. PIM architectures offer a potential solution by performing computations directly within or near memory, reducing data movement. This work is a first look at exploring PIM's potential to accelerate spectral library searches by integrating hardware and algorithm co-design across the MS data analysis pipeline.

We make the following contributions:

- Adaptation of an approximate sum of absolute differences algorithm for PIM-based spectral matching.
- A preliminary approach to integrate PIM into mass spectrometry workflows.
- Early performance assessments via PIM simulation.

II. Background

DRAM and Processing-in-memory (PIM). Modern applications, particularly in fields such as bioinformatics, data analytics, and machine learning, are handling increasingly large amounts of data. This trend has exposed limitations in compute-centric architectures. In order to perform computations on the data stored in memory, data must be transferred between the memory and processor over a narrow memory channel, e.g. 64-bit wide channel for conventional double data rate (DDR) DRAM. For workloads with large data volumes, frequent data movement bottlenecks performance and incurs large energy and latency costs [1].

This work was supported in part by the Jump 2.0 PRISM center, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

One potential solution to the data movement problem is processing-in-memory (PIM) (Fig. 1). The two main approaches to PIM are processing-near-memory (PNM) and processing-using-memory (PUM). PNM architecture employs 3D-stacked memory with a logic layer, which takes advantage of the high bandwidth communication over vertical interconnects between layers to enable PIM. PUM uses the inherent circuit-level properties of memory cells for computation within memory arrays.

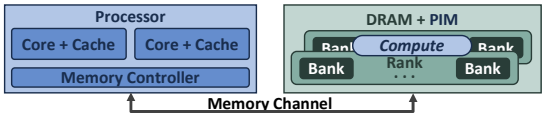

Fig. 1: Processor and DRAM+PIM architecture alleviates data movement bottleneck by adding compute units to DRAM.

Computer systems today commonly use DRAM for main memory. Multi-bank (i.e. multiple sets of independent memory array) DRAM architecture offers high internal bandwidth and parallelism. DRAM-based PIM can benefit from simultaneous access to multiple DRAM arrays for parallel data processing. In this work, we explore how massive parallelism offered by DRAM-based PIM architectures can help the performance of bioinformatic applications.

Mass Spectrometry (MS) and Spectral Library Search. Mass spectrometry (MS) is a method to analyze the proteome of protein and peptide samples. A mass spectrum is represented by a plot of mass-to-charge ratio (m/z) on the y-axis to ion signal intensity on the x-axis. The data for a spectrum consist of a series of peaks, each represented by a pair of values: the m/z and its corresponding intensity. These peaks represent the fragment ions derived from the peptide. Additionally, each spectrum includes metadata such as the precursor m/z, which is the m/z of the intact peptide ion before fragmentation.

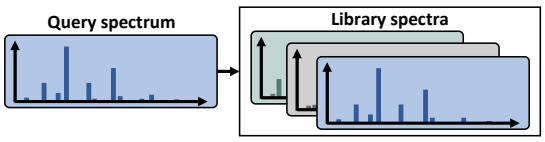

Fig. 2: Spectral library search identifies experimental spectrum by querying it against library spectra.

Spectral library search (Fig. 2) is used to interpret MS/MS data. The search process involves calculating similarity scores

between an experimental spectrum and each library spectrum. With modern MS experiments generating massive data—millions of mass spectra—which must then be matched against libraries of known spectra to identify molecules [2], this comparison becomes expensive. Parallel comparisons of query spectra with many library spectra make it a suitable candidate workload for DRAM-based PIM architectures.

III. Methodology and Implementation

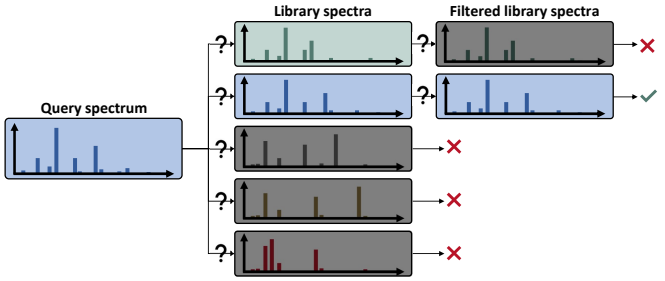

Fig. 3: Pre-filtered spectral matching decreases search space.

Pre-processing and Prefiltering. In the data pre-processing and pre-filtering step (Fig. 3), we begin by discretizing the data similarly to previous work [2], by dividing the m/z range of spectra into bins and summing intensity values of m/z values in the same bin. The data can then be represented by sequences of bin index and intensity pairs. We extend this preprocessing further by exploring approximate computing techniques such as reducing the bit-precision of floating point values and using fixed-point representation for those values. Fixed-point representation simplifies the hardware necessary to process the data and allows for the application to be run on PIM hardware. The low-precision approach aligns well with multiple ppm error tolerance for MS experiments [3], [4].

This strategy enables effective pre-filtering to decrease the search space, not only accelerating standard searches but also facilitating open modification searches where slight mass shifts are allowed. The optimal distribution of these preprocessing steps between the CPU and PIM accelerator is an ongoing area of investigation in our work.

Minimum Sum of Absolute Differences (SAD). We use an adapted minimum SAD algorithm optimized for execution on PIM architectures. Variations of this algorithm are commonly used in the signal processing domain to measure similarity between images. We use SAD to calculate a similarity score between the experimental spectrum $S = (s_1, s_2, \ldots, s_n)$ with each library spectrum $L_i = (\ell_{i1}, \ell_{i2}, \ldots, \ell_{in})$ for $i = 1, 2, \ldots, m$. Assuming m library spectra and n elements in each spectrum, the basic algorithm seeks L_k such that:

$$k = \arg\min_i \sum_{j=1}^{n} |s_j - \ell_{ij}|$$

We implement SAD using PIM API instructions from the PIMeval, a performance and energy simulator for diverse PIM architectures [5]. To increase noise tolerance, we also enable SAD with shifted indices for inexact matching, implemented using the rightward rotation (`pimRotateElementsRight`) instruction:

```
for (int idx=0; idx < subvecLen; idx++) {
  pimSub(obj1, obj2, obj3);
  pimAbs(obj3, obj3);
  for (int i=idx; i+subvecLen-1 < vecLen; i+=subvecLen) {
    pimRedSumRangedInt(obj3, i, i+subvecLen-1, &sumAbsDiff);
    ...
  }
  pimRotateElementsRight(obj2);
}
```

IV. Evaluation

We model subarray-level bit serial PIM for the application on the simulator and conducted initial experiments comparing our PIM-based approach with a traditional CPU implementation for SAD, using our pre-processed data. The parameters used for the PIM device were a single rank DIMM with 8 chips, 16 banks per rank, and 32 subarrays per rank of 8192×8192 cells.

For the current small-scale problem of 11 MB, the CPU implementation completed execution in 175 ms, while the PIM implementation took 1368 ms for the core computation tasks. Our experiments with a 2-rank DIMM configuration for the same problem size revealed that, at this scale, the workload underutilized the available parallelism of the additional hardware resources. These preliminary results highlight an important aspect of PIM architectures: their performance benefits are expected to scale with problem size. While the current small dataset favors the CPU, we anticipate that for larger, more realistic proteomics problem sizes, PIM acceleration will demonstrate advantages in scalability, memory bandwidth utilization, and energy efficiency, being able to benefit from parallel processing and decreases in data movement.

Conclusion and Future Work

We have demonstrated a proof of concept for an end-to-end, hardware-algorithm co-designed pipeline for DRAM-based PIM architecture acceleration of spectral library search, a key workload for mass spectrometry. The next step is to observe the performance from taking better advantage of the massive parallelism in DRAM for more reasonable comparisons and to connect larger volumes of data from real-world workloads to the hardware for a full-system view of PIM accelerated bioinformatics applications.

References

[1] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, "A modern primer on processing in memory," in *Emerging Computing: From Devices to Systems*. Springer, 2023, pp. 171–243.

[2] J. Kang, W. Xu, W. Bittremieux, N. Moshiri, and T. Rosing, "Accelerating open modification spectral library searching on tensor core in high-dimensional space," *Bioinformatics*, vol. 39, no. 7, p. btad404, 2023.

[3] V. R. Pagala, A. A. High, X. Wang, H. Tan, K. Kodali, A. Mishra, K. Kavdia, Y. Xu, Z. Wu, and J. Peng, "Quantitative protein analysis by mass spectrometry," *Protein-protein interactions: Methods and applications*, pp. 281–305, 2015.

[4] J. M. Chick, D. Kolippakkam, D. P. Nusinow, B. Zhai, R. Rad, E. L. Huttlin, and S. P. Gygi, "A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides," *Nature biotechnology*, vol. 33, no. 7, pp. 743–749, 2015.

[5] F. A. Siddique, D. Guo, Z. Fan, M. Gholamrezaei, M. Baradaran, A. Ahmed, H. Abbot, K. Durrer, K. Nandagopal, E. Ermovick *et al.*, "Architectural modeling and benchmarking for digital dram pim."