An In-Storage Processing Architecture with 3D NAND Heterogeneous Integration for Spectra Open Modification Search

Po-Kai Hsu*
Georgia Institute of Technology
pokai.hsu@gatech.edu

Weihong Xu*
University of California, San Diego
wexu@ucsd.edu

Tajana Rosing
University of California, San Diego
tajana@ucsd.edu

Shimeng Yu
Georgia Institute of Technology
shimeng.yu@ece.gatech.edu

ABSTRACT

Spectra open modification search (OMS) is the critical step in mass spectrometry (MS) analysis and proteomics to identify peptides underlying protein samples. However, large-scale spectra OMS is a data-intensive workload that takes hours to days. In this work, we propose a reconfigurable architecture based on 3D NAND ISP with heterogeneous integration to accelerate the mass spectrum data processing. We present two types of encoding designs for optimization. Then we design scalable and reconfigurable 3D NAND ISP tiles to further optimize the performance. The experiments show that the 3D NAND ISP architecture with proper hardware configuration achieves 14.3× to 24.2× speedup over the GPU baseline [10]. The energy consumption is also improved by four orders of magnitude without data movements. The proposed design is an energy-efficient and high-performance ISP solution for the emerging large-scale spectra OMS.

CCS CONCEPTS

• **Hardware → Emerging architectures**; **Memory and dense storage**; **Application specific integrated circuits**; • **Computer systems organization → Parallel architectures**.

KEYWORDS

In-storage processing, 3D NAND ISP, Heterogeneous integration, Mass spectrometry, Open Modification Search, domain-specific acceleration

ACM Reference Format:
Po-Kai Hsu, Weihong Xu, Tajana Rosing, and Shimeng Yu. 2023. An In-Storage Processing Architecture with 3D NAND Heterogeneous Integration for Spectra Open Modification Search. In *The International Symposium on Memory Systems (MEMSYS '23), October 02–05, 2023, Alexandria, VA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3631882.3631896

*Both authors contributed equally to the paper.

MEMSYS '23, October 02–05, 2023, Alexandria, VA, USA

ACM ISBN 979-8-4007-1644-7/23/10.
https://doi.org/10.1145/3631882.3631896

1 INTRODUCTION

Proteomics is a key to understanding the molecular processes of proteins, which are responsible for a variety of activities in cell life. Proteomics scientists use a powerful technique, called mass spectrometry (MS), to recognize and measure peptides and proteins underneath biological samples. Figure 1 illustrates the standard flow to identify peptide sequences contained in protein digestion. First, a method called tandem mass spectrometry (MS/MS) produces a large amount of unknown query spectra data. Second, the key step here is to compare the experimental query spectra against a pre-built spectral reference library with known peptides, using the spectral library searching method [12].

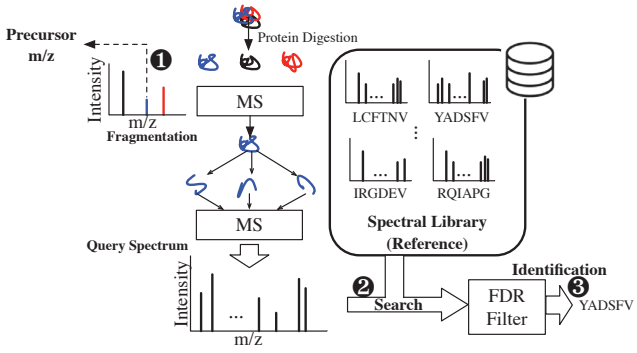


Figure 1: Overview of spectral library searching [9].

The algorithmic challenge of spectral library search is: a large amount of acquired query spectra cannot be directly identified by just using popular similarity metrics (like cosine similarity or inner product) [5]. This is due to the data mismatch between experimental and reference spectra data. The analyzed protein samples may encounter multiple post-translational modifications (PTMs) that modify the inherent mass and MS/MS fragmentation patterns. However, reference spectra in pre-built spectral libraries are mainly unmodified peptides. So more advanced searching algorithm is needed to address PTMs. Open modification searching (OMS) is a promising solution to accurately identify modified spectra [14]. Unlike the standard spectral library search that only queries spectra to reference with a similar precursor mass, OMS accepts reference spectra from a much wider range such that modified query spectra are searched against their unmodified reference variants with different precursor masses.

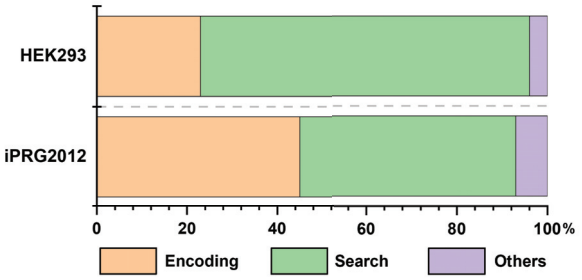
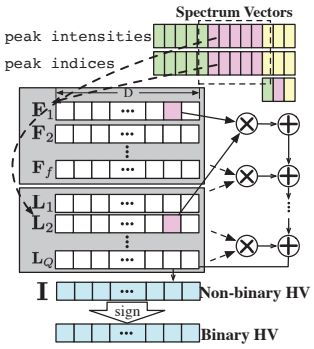


Figure 2: Runtime breakdown of HOMS-TC [10] on GPU.

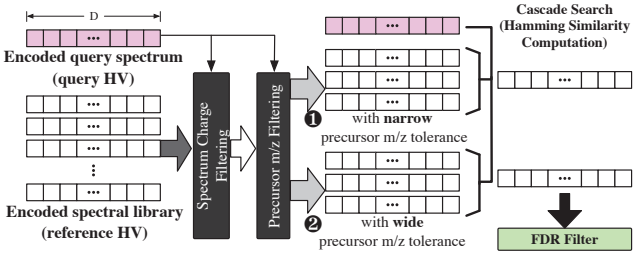
Spectra OMS enables the study of more complex protein interaction in virus-host and proteomics analysis of non-model organisms [8]. However, OMS workloads create three major challenges in terms of algorithm and data analysis acceleration. 1. OMS is a memory-intensive workload that exhibits very low searching speed and efficiency even with careful optimizations [2] since OMS drastically increases the search space. 2. The increasingly available spectra data in public databases [15] promote research development, but the massive spectral libraries created by repository-scale MS data [25] further increases the OMS time from hours to days. For example, UCSD MassIVE contains 5.6 billion spectra, which corresponds to 448TB in size [25].

Several tools have been presented to shorten the OMS time [3, 10, 13]. These tools use advanced nearest-neighbor search algorithms with optimized metrics to boost OMS. Among the state-of-the-art accelerations, HOMS-TC [10] with the aid of hyperdimensional computing (HD) demonstrates the best runtime performance as well as memory efficiency because it leverages the HD technique to simplify the required operations to hardware-friendly Boolean operations while maintaining good searching quality. Although HD-based HOMS-TC significantly speeds up OMS workloads, it still incurs a large memory footprint due to the memory-intensive HD primitives. As shown in Figure 2, the HD encoding and database search dominate the overall runtime even using a NVIDIA RTX 4090 GPU with 1TB/s memory bandwidth.

In-storage procesing (ISP) [17, 21, 22] is considered an effective solution to extend available bandwidth and reduce data movement cost. Meanwhile, the high-density 3D NAND Flash provides a cost-effective solution that allows the storage of spectra data with over GB or TB sizes. In this work, we combine the heterogeneous integration techniques [18] with 3D NAND ISP to develop an architecture to accelerate HD-based OMS workloads in HOMS-TC [10] that shows high data parallelism and energy efficiency. To accommodate the entire reference datasets, several tiles are required, thus offering the reconfigurability of the 3D NAND ISP architecture. We simulate the hardware performance with industry-grade 3D NAND parameters [21] and implement the encoding and search circuits in 7nm FinFET technology node with ASAP7 PDK [6]. The 3D NAND peripheral circuits are extracted from NeuroSim [24]. Our in-house simulator shows the 3D NAND ISP has 14.3× to 24.2× speedup versus the HOMS-TC. The energy efficientcy is also improved by four orders of magnitude without massive data movements.



(a) Encoding.



(b) Hamming similarity search.

Figure 3: Two major steps: (a) encoding and (b) search in HD-based OMS [9]. The encoding step converts spectra peaks into hypervectors. The search step uses Hamming similarity to efficiently find the matched peptides.

2 BACKGROUND ON MS AND ISP

2.1 HD-based Spectra Open Modification Search

Spectra data contain the mass-to-charge ratio (m/z) and ion signal intensity of proteins. We call them peak intensities and peak indices, respectively. Hyperdimensional computing-based (HD-based) OMS improves the efficiency of the conventional spectra OMS pipeline (Figure 1) in two aspects: 1. encoding and 2. Hamming similarity search. In this work, we use the similar HD-based OMS in [9, 10] as the OMS algorithms.

HD Encoding for Spectra. Figure 3 shows the encoding step that transforms the raw spectra data into *hyperdimensional space*, where the spectra are expressed as binary vectors with high dimension, called *hypervectors (HVs)*. To model the peak shifts and intensity changes due to PTMs, HD encoding [9, 10] considers both *spatial locality* (for peak shift) and *value locality* (for peak intensity change). Each index in the spectrum vector is assigned with the associative *position HV* $\mathbf{F}$ such that $\mathbf{F}_i$ corresponds to index i, and $\mathbf{F} \in \{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_f\}$, where f denotes the spectrum vector dimension. Likewise, *level HVs* $\mathbf{L}$ are utilized to model the intensity values in each index. The intensity values are quantized to Q levels and $\mathbf{L}_i$ is assigned to the associative level i where $i \in [0, Q)$.

With the two sets of encoding HVs, namely $\mathbf{F}$ and $\mathbf{L}$, the preprocessed spectrum vector with multiple pairs of peak intensities and

indices are encoded into the HV $\mathbf{I}$ format as:

$$\mathbf{I} = \sum_{(i,j)\in\mathbb{P}} \mathbf{F}_i \odot \mathbf{L}_j, \tag{1}$$

where $\mathbb{P}$ denotes all pairs of peak intensities and indices represent the element-wise multiplication. Note that the resulting aggregated HV $\mathbf{I}$ is non-binary HV. We binarize it for better computation and memory efficiency.

Hamming Similarity Search. After the encoding step, HD-based OMS leverages *Hamming similarity search* to identify the reference peptides in HV format most matched to the query HV. Specifically, *Hamming similarity* is adopted as the search metric. Therefore, the search step requires to compute the Hamming similarity between query and reference HVs. Each spectrum has its own spectrum charge (+2, +3, ...) and precursor *m/z* value. In addition to Hamming similarity, the matched reference HVs also need to satisfy other constraints including the spectrum charge and precursor *m/z* condition. The final search results satisfy both: (1) having the identical spectrum charge as the query and (2) falling into the valid range of precursor *m/z* difference between query and reference.

We apply the *cascade search* [11] to reduce the misidentification rate, where a narrow precursor *m/z* tolerance is firstly used for the standard search and FDR filtration is applied as Figure 3(b)-❶. In the second phase, remaining unidentified spectra are searched using a larger precursor *m/z* tolerance as ❷.

The advantages of HD-based OMS lie in: the binary HV representation instead of the high-precision format in existing OMS tools [3, 13], which only requires simple Hamming similarity operations during OMS. The simplified data format and computations dramatically reduce the circuit complexity for ISP implementation.

2.2 3D NAND In-Storage Processing (ISP)

Large datasets beyond several GB in scale often require Solid State Drives (SSD) to accommodate the entire dataset. While SSDs offer high read-throughput, accessing the entire dataset can still incur significant latency and energy consumption. To address this issue, in-storage-processing (ISP) has been proposed as a promising paradigm [17, 21, 22] to eliminate the overhead caused by data movements. Figure 4 illustrates the configuration of 3D NAND ISP. In this design, an additional set of Analog-to-Digital Converters (ADCs) is integrated into the separated source line (SL) corresponding to each block in the mature 3D NAND Flash configuration. The weight matrix or the reference data is stored in the 3D NAND Flash, while the input vector or the query is sent to the 3D NAND as bit line (BL) voltages. The results of either the vector-matrix multiplication of the input vector and the weight matrix or the dot product of the reference data and the query equal to the summed currents along the sourcelines (SLs). The ADC then converts this current into the digital domain for post-ISP processing. Without the need for GB-level data movements, 3D NAND ISP reduces overall latency and lowers energy consumption. As a result, in-storageprocessing holds great potential for optimizing the performance of systems dealing with large datasets on SSDs.

2.3 Heterogeneous Integration

To further boost the performance, heterogeneous integration techniques are proposed to stack peripheral circuits on top/bottom of the 3D NAND Flash array. Incorporating with Cu-Cu hybrid bonding [19] and CMOS under array (CUA) [20], ISP achieves a compact form factor. CUA enables the overlapping of memory peripherals under the array, reducing the area of a single tier. Meanwhile, the high-density inter-chip Cu-Cu bonding connects the processing elements on the CMOS wafer to the 3D NAND wafer, ensuring seamless integration. The CMOS wafer can be fabricated in an advanced technology node to yield a smaller area and better performance. The combination of CIM with heterogeneous integration [18] offers a compact solution for large-scale data processing with enhanced performance. This approach opens new possibilities for the development of low-power, high-performance, and compact data processing systems applicable to various applications.

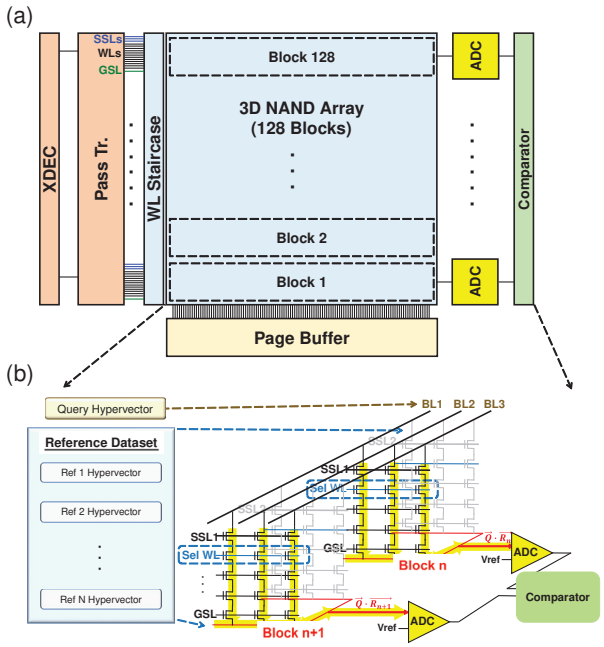


Figure 4: Overview of 3D NAND in-storage processing (ISP) architecture: (a) Configuration of 3D NAND ISP. An additional set of ADCs are deployed after separated SLs, which converts The VMM results or dot product results into digital domain. (b) Data mapping scheme of 3D NAND ISP. Taking OMS for example, the reference dataset is mapped to the 3D NAND array and the query for hamming similarity calculation are sent in the array as BL voltage. The summed currents in the SL represent the dot product results and later sorted after ADCs.

3 PROPOSED 3D NAND ISP ARCHITECTURE

The datasets for mass spectrometry have reference data in the number of million-level. In this work, we propose a reconfigurable architecture based on 3D NAND ISP with heterogeneous integration for mass spectrometry applications. The 3D NAND ISP tile possesses the capability to perform both query encoding and hamming similarity search in HyperOMS. In this section, the architecture of 3D NAND ISP and reconfigurability are discussed.

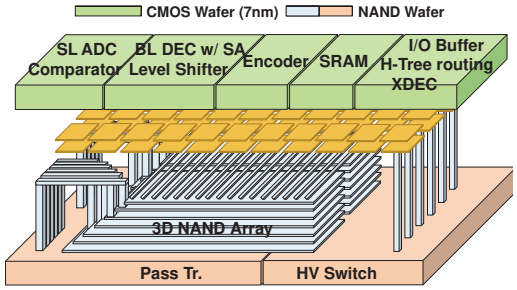


Figure 5: 3D NAND ISP tile with heterogeneous integration. The high-voltage circuits are stacked underneath the 3D NAND array using CUA. The low-voltage circuits and digital circuits are fabricated on a separated CMOS wafer in an advanced technology node and. The 3D NAND wafer and CMOS wafer are bonded using Cu-Cu bonds offering high-bandwidth inter-tier communication.

3.1 3D NAND ISP Tile with Heterogeneous Integration

Figure 5 shows the proposed 3D NAND ISP tile with heterogeneous integration. The peripheral circuits are folded on the top and bottom of the 3D NAND tile. Notably, the high-voltage circuits including word line (WL)/string select line (SSL) switch matrix (SW) and the pass transistors are fabricated underneath the 3D NAND array using CUA approach with the transistor size equivalent to 65 nm technology to sustain high-voltage program/erase operations of 3D NAND Flash. On the other hand, the low-voltage circuits including digital circuits, buffers, decoders and ADCs are fabricated on a separate CMOS wafer in an advanced 7 nm technology node and later face-to-face bonded on top of the 3D NAND wafer using Cu-Cu hybrid bonding. The inter-tier Cu-Cu bonding has a tight pitch of 1 μm [23] to guarantee high bandwidth data communication across tiers. With Heterogeneous integration, the 3D NAND ISP can accommodate encoding circuits and search circuits, therefore performing both encoding and OMS in a single compact tile.

3.2 In-Memory Encoding vs. Near-Memory Encoding

The hardware implementation of XOR encoding can also be incorporated in an in-storage fashion. Unlike the previous ISP approach for dot products on SLs, the in-memory encoding performs bit-wise dot products on each BL. Figure 6 illustrates both the near-memory and in-memory encoding hardware designs. The near-memory encoding method deploys a set of XOR gates after sense amplifiers (SA) in the page buffer. The *position HVs* are read from 3D NAND Flash and fed into the XOR gates alongside cached *level HVs*. On the other hand, in the in-memory encoding design, the *position HVs* are also stored in the 3D NAND array, while in need of storing *position* $\overline{HVs}$ and the *level HVs* are sent in as the BL voltages. The XOR operation can be replaced by the OR operation of two bit-wise dot products as:

$$\mathbf{A} \oplus \mathbf{B} = \left(\bar{\mathbf{A}} \cdot \mathbf{B}\right) \vee \left(\mathbf{A} \cdot \bar{\mathbf{B}}\right). \tag{2}$$

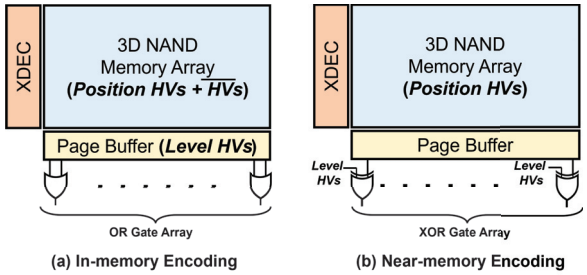


Figure 6: Block diagrams of in-memory encoding and near-memory encoding: (a) In-memory encoding. The *position HVs* and *position* $\overline{HVs}$ are stored in the 3D NAND array. The XOR encoding is achieved by the OR result of two dot products. (b) Near-memory encoding. The *position HVs* are read from the 3D NAND array and complete the XOR encoding with the cached *level HVs*.

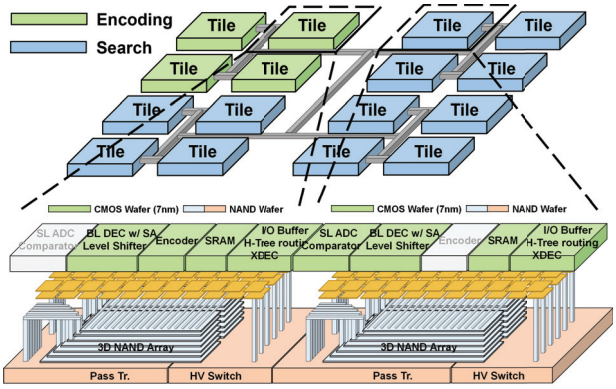


Figure 7: Reconfigurable 3D NAND ISP architecture. The tile performs encoding and search operation. Combining several tiles with H-tree routing provides flexibility to assigned encoding or search to the specified tile for optimization.

Integrating a set of AND gates after two sense amplifiers, the in-memory requires less logic area with respect to the simplicity of the OR gate compared to the XOR gate. The tradeoff will be discussed in the **Evaluation** section.

3.3 Reconfigurability

Since the 3D NAND ISP tile performs encoding and search, multiple tiles can be partitioned for specific tasks, e.g., encoding and search tiles. The versatility offers the reconfigurability for the chip to accelerate specified tasks with optimized tile designs. Figure 7 demonstrates the reconfigurable architecture of the 3D NAND ISP tiles. The tiles communicate through H-tree routing on the top CMOS tier with memory controllers. This H-tree routing offers inter-tile communications including tile-to-tile data transmission and broadcasting. The reconfigurable architecture design provides a design space for optimization when dealing with various datasets with different parameters.

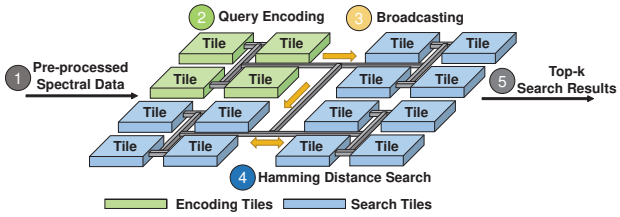


Figure 8: Data flow in the 3D NAND ISP architecture. The preprocessed spectra are fetched and encoded by the encoding tiles. The encoded query is broadcast to the search tiles for the Hamming similarity search in parallel. Finally, the sorted top-k results are sent out.

Table 1: Datasets and spectrum preprocessing configurtions.

	Dataset	
Parameter Name	**iPRG2012 [4]**	**HEK293 [5]**
Max peaks in spectra	50	
Min / max m/z	101 / 1500	
Bin size	0.05	0.04
Precursor m/z tolerance (narrow)	20ppm	5ppm
Precursor m/z tolerance (wide)	500Da	500Da

3.4 Data Flow

Figure 8 illustrates the data flow of the architecture. First, the pre-processed spectral data is fetched through the IO sequentially. The specified encoding tiles encode the pre-processed spectral data into query hypervectors, which are subsequently broadcasted to the search tiles for simultaneous parallel searching. Finally, the hamming similarities are sorted after exploring all the search spaces, and the top-k results are sent out serially through the IO interface.

4 EVALUATION

4.1 Methodology

Datasets. We use two real-world datasets, including: 1. small-scale iPRG2012 dataset [4] (total spectra: 15, 867) as query while yeast spectral dataset [16] with the human HCD spectral library (total spectra: 1, 162, 392) as reference. 2. large-scale HEK293 (Human Embryonic Kidney 293) dataset [5] (total spectra per query: 46, 665 on average) as query while the human spectral library [1, 26] (total spectra: 2, 992, 672) as reference. The query and reference spectra follow the preprocessing flow of existing works [2, 3, 10]. The preprocessing configurations for query and reference spectra are listed in Table 1. The low-quality spectra with less than ten peaks and a 250 m/z mass range or peaks within a 0.05 m/z window around the precursor m/z were removed. All MS data, spectral libraries, preprocessed spectra, and identification results are available on the MassIVE repository with the dataset identifier MSV000091183.

Benchmarking. The evaluation of software baselines is run on Intel i7-11700K CPU with 64GB of RAM, and NVIDIA Geforce RTX 4090 with 24GB of VRAM. We measure the energy consumption of the CPU and GPU using Intel Power Gadget and `nvidia-smi`, respectively. We count the number of identifications to compare

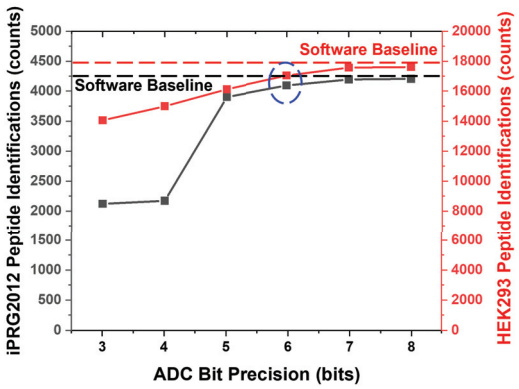


Figure 9: Impact of ADC precision on the OMS search quality in terms of identified peptides.

the search quality. All search results are evaluated at fixed 1% FDR threshold, using Pyteomics [7].

Hardware Modeling. The hardware parameters of the proposed 3D NAND are listed in Table 2. The HD encoder and search circuits are implemented using *Verilog* and synthesized on ASAP 7nm PDK[6]. The peripheral circuits of the 3D NAND array are extracted from NeuroSim [24]. The clock frequency is set to 1GHz. To estimate the performance and energy efficiency of proposed ISP designs, we develop an in-house simulator to run the trace extracted from the HOMS-TC [10] software.

Table 2: Hardware Simulation Parameters

	Paramters	Values
Advanced CMOS Tier	Technology	7 nm FinFET Process
	VDD1	0.7 V
	ADC Type	6-bit SAR ADC
	Encoder Dimension	8192
3D NAND Physical Parameter[21]	Equivalent Feature Size F	13 nm
	SSL Pitch	220 nm
	BL Pitch	100 nm
	No. of WL	32
	No. of SSL	16
	No. of BL	1/2/4/8 KB
	No. of Block	128
	Tile Size	0.379/0.757/1.51/3.03 mm^2
	WL Staircase Pitch	500 nm
3D NAND Electrical Parameter[21]	WL Read Voltage	1V/4.5 V (V_{select}/V_{pass})
	SSL Read Voltage	4.5 V (activated)
	BL Read Voltage	0.2V
	I_{on}/I_{off}	2 nA/1 pA
CMOS under Array	Technology	65 nm Process
	VDD2	1V

4.2 Performance and Energy Evaluation

ADC precision. To simulate the performance, the ADC precision for the 3D NAND ISP is needed to be determined. ADC introduces additional quantization errors, which degrades the accuracy. Figure 9 demonstrates the impact of ADC precision on the OMS search quality. The quantization error is negligible when ADC is 6-bit. Therefore, we design the ADCs with 6-bit SAR ADC.

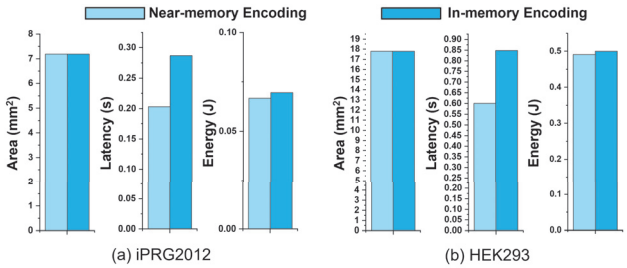


Figure 10: Hardware simulation results of in-memory encoding versus near-memory encoding. Note that BL number is 8192.

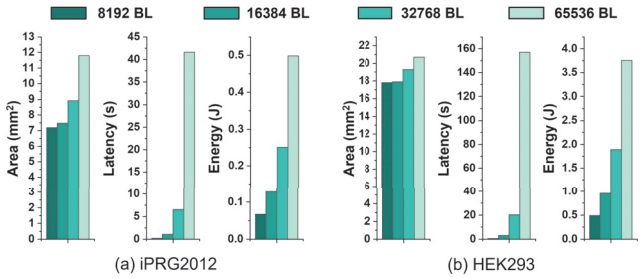


Figure 11: Hardware simulation results of various page sizes (1KB/2KB/4KB/8KB) for near-memory encoding implementation.

In-memory encoding vs near-memory encoding. For the 3D NAND ISP hardware evaluation, we first compare the performance of the two hardware implementation methods for encoding. Figure 10 shows the simulation results of in-memory encoding and near-memory encoding. Note that the BL number is set to 1KB (8192) for fair comparison. Although in-memory encoding can reduce the circuit complexity, the doubled read operations for *position HVs* yield longer latency and larger energy consumption for the specific XOR encoding approach. In-memory encoding will outperform near-memory encoding in the more complex encoding methods. Later simulations are based on near-memory encoding.

Page size scaling. The latency and energy consumption of a 3D NAND memory array is dominated by the WL charging/discharging. Therefore, a sizable page offers a degree of freedom to further optimize the performance. Figure 11 shows the hardware simulation results of various page sizes, i.e., numbers of BL. We selectively simulate 1KB(8192), 2KB(16384), 4KB(32768) and 8KB (65536). With respect to the dimension of hypervectors is 8192, the minimum number of BL is set to 8192 to avoid additional partial sum overhead. The simulation results show a larger number of BL yields worse performance. This is because the latency and energy consumption of WL operations are scaled accordingly. We propose to design the 3D NAND ISP with a minimum page size that equals the dimension of hypervectors for agile operations.

Tile scaling. The reconfigurable design also provides the scalability for further speedup. Figure 12 shows the hardware simulation results of scaled tile numbers. As the number of tile scales, the latency is decreased. However, the scaling of latency is not inversely linear due to the digital processing overhead. We propose to scale the tile number by 2× to obtain an optimized result with a reasonable area of 14.4 and 35.6 mm^2 for iPRG2012 and HEK293, respectively.

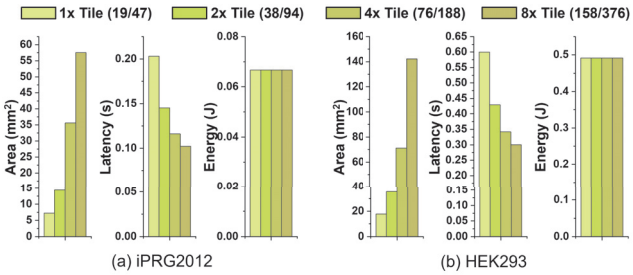


Figure 12: Hardware simulation results of scaled tile numbers for near-memory encoding implementation. Note that BL number is 8192.

Table 3: Speedup over the state-of-the-art OMS library on GPU, HOMS-TC [10]. The HEK293 runtime is the average runtime for each query file.

Workload	Spectra OMS	
Dataset	iPRG2012	HEK293
HOMS-TC [10]	2.08s (1×)	10.4s (1×)
This work	0.145s (14.3×)	0.429s (24.2×)

Speedup versus GPU. With the optimized configuration of 3D NAND ISP, we compare the performance versus CPU and GPU. Table 3 compares the latency for HOMS-TC which accelerates HyperOMS on GPU and HyperOMS on 3D NAND ISP. The proposed 3D NAND ISP has 14.3× and 24.2× speedup on respective datasets. The simulated energy consumptions are 0.067 J and 0.491 J. Considering the average power of GPU 450 W, 3D NAND ISP improves the energy efficiency by four orders of magnitude.

5 CONCLUSION

In this work, we propose the 3D NAND ISP architecture to accelerate memory-intensive spectral open modification search (OMS) workloads. We also present two types of encoding design and determine the near-memory encoding for the state-of-the-art HD-based OMS algorithm [9, 10]. The proposed 3D NAND ISP provides reconfigurability and scalability for further optimization. Without the need to move massive data from SSD and memory, the energy consumption is significantly reduced by four orders of magnitude and 14.3× to 24.2× speedup is achieved over the GPU baseline [10]. Our design is an energy-efficient and high-performance ISP solution for the emerging large-scale spectra OMS.

ACKNOWLEDGMENTS

This work is supported by PRISM, one of the SRC/DARPA JUMP 2.0 centers. The authors thank Macronix, Taiwan, for providing the technical specifications for the 3D NAND prototype.

REFERENCES

[1] 2022. *MassIVE-KB.* https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp

[2] Wout Bittremieux et al. 2018. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *Journal of Proteome Research* 17, 10 (Sept. 2018), 3463–3474.

[3] Wout Bittremieux et al. 2019. Extremely Fast and Accurate Open Modification Spectral Library Searching of High-Resolution Mass Spectra Using Feature Hashing and Graphics Processing Units. *Journal of Proteome Research* 18, 10 (Aug. 2019), 3792–3799.

[4] Robert J Chalkley et al. 2014. Proteome Informatics Research Group (iPRG)_2012: A Study on Detecting Modified Peptides in a Complex Mixture. *Molecular & Cellular Proteomics* 13, 1 (Jan. 2014), 360–371.

[5] Joel M Chick et al. 2015. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nature Biotechnology* 33, 7 (June 2015), 743–749.

[6] Lawrence T. Clark et al. 2016. ASAP7: A 7-nm finFET predictive process design kit. *Microelectronics Journal* 53 (2016), 105–115.

[7] Anton A. Goloborodko et al. 2013. Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. *Journal of The American Society for Mass Spectrometry* 24, 2 (Jan. 2013), 301–304.

[8] Michelle Heck et al. 2020. Proteomics in Non-model Organisms: A New Analytical Frontier. *Journal of Proteome Research* 19, 9 (2020), 3595–3606.

[9] Jaeyoung Kang et al. 2022. Massively Parallel Open Modification Spectral Library Searching with Hyperdimensional Computing. In *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*. 536–537.

[10] Jaeyoung Kang et al. 2023. Accelerating open modification spectral library searching on tensor core in high-dimensional space. *Bioinformatics* 39, 7 (2023).

[11] Attila Kertesz-Farkas et al. 2015. Tandem Mass Spectrum Identification via Cascaded Search. *Journal of Proteome Research* 14, 8 (2015), 3027–3038.

[12] Henry Lam. 2011. Building and Searching Tandem Mass Spectral Libraries for Peptide Identification. *Molecular & Cellular Proteomics* 10, 12 (Dec. 2011), R111.008565–R111.008565.

[13] Henry Lam et al. 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. *PROTEOMICS* 7, 5 (March 2007), 655–667.

[14] Seungjin Na and Eunok Paek. 2015. Software Eyes for Protein Post-Translational Modifications. *Mass Spectrometry Reviews* 34, 2 (April 2015), 133–147.

[15] Yasset Perez-Riverol et al. 2022. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research* 50, D1 (2022), 543–552.

[16] Nathalie Selevsek et al. 2015. Reproducible and Consistent Quantification of the Saccharomyces cerevisiae Proteome by SWATH-mass spectrometry. *Molecular & Cellular Proteomics* 14, 3 (March 2015), 739–749.

[17] Wonbo Shim et al. 2021. Architectural Design of 3D NAND Flash Based Compute-in-Memory for Inference Engine. In *Proceedings of the International Symposium on Memory Systems*. 77–85.

[18] Wonbo Shim and Shimeng Yu. 2021. Technological Design of 3D NAND-Based Compute-in-Memory Architecture for GB-Scale Deep Neural Network. *IEEE Electron Device Letters* 42, 2 (2021), 160–163.

[19] TechInsights. 2020. YMTC 64L TLC 3D NAND Product Brief. https://www.techinsights.com/ko/node/31226.

[20] Christian Caillat *et al.* 2017. 3DNAND GIDL-Assisted Body Biasing for Erase Enabling CMOS under Array (CUA) Architecture. In *IEEE International Memory Workshop (IMW)*. 1–4. https://doi.org/10.1109/IMW.2017.7939067

[21] Hang-Ting Lue *et al.* 2019. Optimal Design Methods to Transform 3D NAND Flash into a High-Density, High-Bandwidth and Low-Power Nonvolatile Computing in Memory (nvCIM) Accelerator for Deep-Learning Neural Networks (DNN). In *IEEE International Electron Devices Meeting (IEDM)*. 38.1.1–38.1.4.

[22] Han-Wen Hu *et al.* 2022. ICE: An Intelligent Cognition Engine with 3D NAND-based In-Memory Computing for Vector Similarity Search Acceleration. In *IEEE/ACM MICRO*. 763–783.

[23] R. Chen *et al.* 2020. 3D-optimized SRAM Macro Design and Application to Memory-on-Logic 3D-IC at Advanced Nodes. In *IEEE International Electron Devices Meeting (IEDM)*. 15.2.1–15.2.4.

[24] Xiaochen Peng *et al.* 2019. DNN+NeuroSim: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators with Versatile Device Technologies. In *IEEE International Electron Devices Meeting (IEDM)*. 32.5.1–32.5.4.

[25] UCSD. 2022. MassIVE: Mass Spectrometry Interactive Virtual Environment. https://massive.ucsd.edu/.

[26] Mingxun Wang et al. 2018. Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* 7, 4 (Oct. 2018), 412–421.e5.