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

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

Weihong Xu
University of California, San Diego
wexu@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, HyperOMS, domain-specific acceleration

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].

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.

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. 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 OMS 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 processing (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 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 efficiency is also improved by four orders of magnitude without massive data movements.

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 F such that Fj corresponds to index i, and F ∈ {F1, F2, . . . , Ff}, where f denotes the spectrum vector dimension. Likewise, level HVs L are utilized to model the intensity values in each index. The intensity values are quantized to Q levels and Lj is assigned to the associative level i where i ∈ [0, Q).

With the two sets of encoding HVs, namely F and L, the preprocessed spectrum vector with multiple pairs of peak intensities and indices are encoded into the HV I format as:

\[ I = \sum_{(i,j) \in P} F_j \odot L_j, \]

where P denotes all pairs of peak intensities and indices represent the element-wise multiplication. Note that the resulting aggregated HV 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)-1. In the second phase, remaining unidentified spectra are searched using a larger precursor m/z tolerance as 2.

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-storage-processing 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 CM 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.

3 Proposed 3D NAND ISP Architecture

The datasets for large scale 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.

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,
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.

Figure 6: Block diagrams of in-memory encoding and near-memory encoding: (a) In-memory encoding. The position HVs and position 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.

the position HVs are also stored in the 3D NAND array, while in need of storing position HVs and 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:

\[ A \oplus B = (\bar{A} \cdot B) \lor (A \cdot \bar{B}). \] (2)

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.

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.

### Table 1: Datasets and spectrum preprocessing configurations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max peaks in spectra</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>Min / max m/z</td>
<td>101 / 1500</td>
<td>101 / 1500</td>
<td>101 / 1500</td>
</tr>
<tr>
<td>Bin size</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Precursor m/z tolerance (narrow)</td>
<td>20ppm</td>
<td>5ppm</td>
<td>5ppm</td>
</tr>
<tr>
<td>Precursor m/z tolerance (wide)</td>
<td>500Da</td>
<td>500Da</td>
<td>500Da</td>
</tr>
</tbody>
</table>

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
Table 2: Hardware Simulation Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced CMOS Tier</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>7 nm FinFET Process</td>
</tr>
<tr>
<td>VDD1</td>
<td>0.7 V</td>
</tr>
<tr>
<td>ADC Type</td>
<td>6-bit SAR ADC</td>
</tr>
<tr>
<td>Encoder Dimension</td>
<td>8192</td>
</tr>
<tr>
<td>3D NAND Physical Parameter[21]</td>
<td></td>
</tr>
<tr>
<td>Equivalent Feature Size F</td>
<td>15 nm</td>
</tr>
<tr>
<td>SSL Pitch</td>
<td>220 nm</td>
</tr>
<tr>
<td>BL Pitch</td>
<td>100 nm</td>
</tr>
<tr>
<td>No. of WL</td>
<td>32</td>
</tr>
<tr>
<td>No. of SSL</td>
<td>16</td>
</tr>
<tr>
<td>No. of BL</td>
<td>1/2/4/8 KB</td>
</tr>
<tr>
<td>No. of Block</td>
<td>128</td>
</tr>
<tr>
<td>Tile Size</td>
<td>0.379/0.575/1.53/3.03 mm²</td>
</tr>
<tr>
<td>WL Staircase Pitch</td>
<td>500 nm</td>
</tr>
<tr>
<td>3D NAND Electrical Parameter[21]</td>
<td></td>
</tr>
<tr>
<td>WE Read Voltage</td>
<td>(Vselect/Vpass)</td>
</tr>
<tr>
<td>SSL Read Voltage</td>
<td>4.5 V (activated)</td>
</tr>
<tr>
<td>BL Read Voltage</td>
<td>0.2V</td>
</tr>
<tr>
<td>I&lt;sub&gt;ON&lt;/sub&gt;/I&lt;sub&gt;OFF&lt;/sub&gt;</td>
<td>2 nA/1 pA</td>
</tr>
<tr>
<td>CMOS under Array</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>65 nm FinFET Process</td>
</tr>
<tr>
<td>VDD2</td>
<td>0.7 V</td>
</tr>
</tbody>
</table>

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.

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 2x to obtain an optimized result with a reasonable area of 14.4 and 35.6 mm² for iPRG2012 and HEK293, respectively.

Speedup versus GPU. With the optimized configuration of 3D NAND ISP, we compare the performance versus CPU and GPU.
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 14.3x to 24.2x 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.

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.

<table>
<thead>
<tr>
<th>Workload</th>
<th>Dataset</th>
<th>Spectra OMS</th>
<th>HOMS-TC [10]</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMS-TC</td>
<td>[4]</td>
<td>2.08s (1.94)</td>
<td>13.84s (1.94)</td>
</tr>
<tr>
<td>This work</td>
<td>[2]</td>
<td>0.145s (14.3x)</td>
<td>0.429s (24.2x)</td>
</tr>
</tbody>
</table>

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 14.3x and 24.2x 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.

References