


# A framework for high-throughput sequence alignment using real processing-in-memory systems

**Journal Article****Author(s):**

Diab, Safaa; Nassereldine, Amir; Alser, Mohammed; [Gómez Luna, Juan](#) ; Mutlu, Onur; El Hajj, Izzat

**Publication date:**

2023-05

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000613945>

**Rights / license:**

[Creative Commons Attribution 4.0 International](#)

**Originally published in:**

Bioinformatics 39(5), <https://doi.org/10.1093/bioinformatics/btad155>

## Sequence analysis

# A framework for high-throughput sequence alignment using real processing-in-memory systems

Safaa Diab <sup>1,\*</sup>, Amir Nassereldine<sup>1</sup>, Mohammed Alser<sup>2</sup>, Juan Gómez Luna<sup>2</sup>, Onur Mutlu<sup>2,\*</sup>, Izzat El Hajj<sup>1,\*</sup>

<sup>1</sup>Department of Computer Science, American University of Beirut, Riad El-Solh, Beirut 1107 2020, Lebanon

<sup>2</sup>Department of Information Technology and Electrical Engineering, ETH Zürich, Gloriastrasse 35, Zürich 8092, Switzerland

\*Corresponding authors. Department of Computer Science, American University of Beirut, Beirut, Lebanon. E-mail: syd04@mail.aub.edu (S.D.); Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland. E-mail: omutlu@ethz.ch (O.M.); Department of Computer Science, American University of Beirut, Beirut, Lebanon. E-mail: izzat.elhajj@aub.edu.lb (I.E.H.)

Associate Editor: Pier Luigi Martelli

Received 3 August 2022; revised 24 February 2023; accepted 25 March 2023

## Abstract

**Motivation:** Sequence alignment is a memory bound computation whose performance in modern systems is limited by the memory bandwidth bottleneck. Processing-in-memory (PIM) architectures alleviate this bottleneck by providing the memory with computing competencies. We propose Alignment-in-Memory (AIM), a framework for high-throughput sequence alignment using PIM, and evaluate it on UPMEM, the first publicly available general-purpose programmable PIM system.

**Results:** Our evaluation shows that a real PIM system can substantially outperform server-grade multi-threaded CPU systems running at full-scale when performing sequence alignment for a variety of algorithms, read lengths, and edit distance thresholds. We hope that our findings inspire more work on creating and accelerating bioinformatics algorithms for such real PIM systems.

**Availability and implementation:** Our code is available at <https://github.com/safaad/aim>.

## 1 Introduction

One of the most fundamental computational steps in most genomic analyses is *sequence alignment*. This step is formulated as an *approximate string matching* (ASM) problem (Navarro 2001), which typically uses a dynamic programming (DP) algorithm to optimally calculate the type, location, and number of differences in one of the two given genomic sequences. Such sequence alignment information is typically needed for DNA sequence alignment, gene expression analysis, taxonomy profiling of a multi-species metagenomic sample, rapid surveillance of disease outbreaks, and many other important genomic applications.

DP-based alignment algorithms, such as Needleman–Wunsch (NW) (Needleman and Wunsch 1970) and Smith–Waterman–Gotoh (SWG) (Gotoh 1982), are computationally expensive as they have quadratic time and space complexity (i.e.  $O(n^2)$  for a sequence length of  $n$ ). It is mathematically proven that subquadratic alignment algorithm cannot exist (Backurs and Indyk 2015). Recent attempts for improving sequence alignment tend to follow one of three key directions: (i) accelerating the DP algorithms using hardware accelerators, (ii) accelerating the DP algorithms using heuristics and limited functionality, and (iii) reducing the workload for alignment by filtering out highly dissimilar sequences using pre-alignment filtering algorithms. Comprehensive surveys have been done on these existing attempts (Alser et al. 2020a,c, 2022).

The first direction accelerates *exact* sequence alignment using existing hardware devices, such as SIMD-capable multi-core CPUs, GPUs, and FPGAs, or using to-be-manufactured devices, such as application-specific integrated circuits. One of the most recent alignment algorithms, the wavefront algorithm (WFA) (Marco-Sola et al. 2020a,b), indeed benefits from acceleration via SIMD (Marco-Sola et al. 2020a,b, 2022), GPUs (Aguado-Puig et al. 2022), and FPGAs (Haghi et al. 2021). Parasail (Daily 2016), BWA-MEM2 (Vasimuddin et al. 2019), and mm2-fast (Kalikar et al. 2022) all exploit SIMD-capable and multi-core CPUs to accelerate sequence alignment in read mapping. SillaX (Fujiki et al. 2018) provides an order of magnitude of acceleration through specialization, but requires fabricating their architecture designs into real hardware, which is costly and performed at a semiconductor fabrication facility.

The second direction includes limiting the functionality of sequence alignment to performing only edit distance calculation, as in Edlib (Šošić and Šikić 2017), or limiting the number of calculated entries in the DP table, as in windowing/tiling the DP table (Arlazarov et al. 1970, Rizk and Lavenier 2010, Turakhia et al. 2018) and the X-drop algorithm (Zhang et al. 2000) implemented in one of the versions of KSW2 (Li 2018). This second direction is not mutually exclusive from the first and can also benefit from

acceleration via SIMD (Li 2018), GPUs (Ahmed et al. 2020), and FPGAs (Banerjee et al. 2018).

The third direction is to *early* and *quickly* detect any two dissimilar genomic sequences, which differ by more than a user-defined edit distance threshold, and exclude them from being aligned as their alignment result is not useful. Pre-alignment filtering usually saves a significant amount of time by avoiding DP-based alignment *without* sacrificing alignment accuracy or limiting the algorithm functionality as demonstrated when using even basic CPU implementations (Rasmussen et al. 2006, Rizk and Lavenier 2010, Xin et al. 2013, Alser et al. 2020b). Pre-alignment filtering can also benefit from hardware acceleration (Xin et al. 2015, Alser et al. 2017a,b, Kim et al. 2018, Alser et al. 2019, 2020b, Cali et al. 2020).

Regardless of which of these three directions are followed to improve performance, sequence alignment remains a fundamentally memory-bounded computation with a low data reuse ratio (Gupta et al. 2019, Kaplan et al. 2020, Cali et al. 2020, Lavenier et al. 2020). Sequence alignment implementations suffer from wasted execution cycles due to the memory bandwidth bottleneck faced when moving data between the memory units and the computing units (e.g. CPUs, FPGAs, GPUs). This bottleneck exists because of the large disparity in performance between the compute units and memory units in modern computing systems. Following Moore's law, the number of transistors in processors has been doubling about every 2 years, leading to an exponential increase in the power of the processor cores (Moore 1998). However, memory performance did not scale comparably which has made the cost of transferring data between the main memory and the CPU more expensive than the computations to be performed by a CPU instruction on the data (Mutlu et al. 2019, 2020).

Processing-in-memory (PIM) architectures alleviate the data movement bottleneck of modern computing systems by providing the memory with computing competencies (Hwu et al. 2017, Mutlu et al. 2019, 2020, Ankit et al. 2019, 2020, Huang et al. 2021, Hajinazar et al. 2021a,b, Ferreira et al. 2022, Mansouri Ghiasi et al. 2022). PIM has been used to improve the performance of a wide variety of memory-bound computations, including sequence alignment as is the case in RAPID (Gupta et al. 2019), BioSEAL (Kaplan et al. 2020), GenASM (Cali et al. 2020), and SeGraM (Cali et al. 2022). However, these PIM-based sequence alignment solutions rely on emerging technologies that require either major changes to existing hardware or fabricating new hardware chips that are specially designed for the subject algorithm. These limitations pose a critical barrier to the adoption of PIM in sequence alignment.

UPMEM is the *first* publicly available programmable PIM system in the market (Devaux 2019). The UPMEM architecture integrates conventional DRAM arrays and general-purpose cores called DPUs into the same chip. This architecture allows computations to be performed near where the data resides, which reduces the latency imposed by data movement. UPMEM systems have been used to accelerate memory-bound applications such as database index search, compression/decompression, image reconstruction, genomics, and many others (Lavenier et al. 2016a,b,c, Church et al. 2011, Diab et al. 2022, Zois et al. 2018, Lavenier et al. 2020, Nider et al. 2021, Gómez-Luna et al. 2021a,b, 2022, Giannoula et al. 2022).

Our goal is to evaluate the suitability of real PIM systems for accelerating sequence alignment algorithms. To this end, we introduce Alignment-in-Memory (AIM), a framework for PIM-based sequence alignment that targets the UPMEM system. AIM dispatches a large number of sequence pairs across different memory modules and aligns each pair using compute cores within the memory module where the pair resides. AIM supports multiple alignment algorithms including NW, SWG, GenASM, WFA, and WFA adaptive. Each algorithm has alternate implementations that manage the UPMEM memory hierarchy differently and are suitable for different read lengths.

We evaluate AIM on a real UPMEM system and compare the throughput it can achieve with that achieved by server-grade multi-threaded CPU systems running at full scale. Our evaluation shows that a real PIM system can substantially outperform CPU systems for a wide variety of algorithms, read lengths, and edit distance

thresholds. For example, for WFA adaptive, the state-of-the-art sequence alignment algorithm, AIM achieves a speedup of up to  $2.56\times$  when the data transfer time between the CPU and DPUs is included, and up to  $28.14\times$  when that data transfer time is not included. These speedups to sequence alignment can translate into substantial performance improvements for widely used bioinformatics tools such as minimap2, where alignment can consume up to 76% of the time (Kalikar et al. 2022), or BWA-MEM, where alignment can consume up to 47.2% of the time (Vasimuddin et al. 2019). Our results demonstrate that emerging real PIM systems are promising platforms for accelerating sequence alignment. We hope that our findings inspire more work on creating and accelerating bioinformatics algorithms for real PIM systems.

## 2 System and methods

In this section, we provide an overview of PIM and the UPMEM PIM architecture (Section 2.1). We then describe the overall workflow of our PIM-based sequence alignment framework (Section 2.2), how we manage data within the UPMEM PIM memory hierarchy (Section 2.3), and how each of the different sequence alignment algorithms are supported within our framework (Section 2.4).

### 2.1 PIM and the UPMEM PIM architecture

Figure 1 compares the organization of conventional CPU processing systems to PIM systems. In conventional systems, illustrated in Fig. 1a, data resides in dynamic random access memory (DRAM) which is typically organized into multiple DRAM banks. This data is transferred to the CPU cores where computations are performed on the data. If the same data is reused for many computations, the cost of moving that data from memory to the CPU cores is amortized. However, if the data is not reused, the cost of moving the data is much higher than the cost of the computations on that data. In this case, the CPU cores will be idle most of the time waiting for memory accesses to complete, and the movement of data between the CPU cores and the memory becomes a major performance bottleneck.

In a PIM system, illustrated in Fig. 1b, small PIM cores are placed in the memory chip near the memory banks. These PIM cores are typically much less powerful than CPU cores at performing computations; however, they are much faster at accessing data from memory because of their proximity. Hence, if only a few computations are performed on the data, it is cheaper to perform these computations in the PIM cores than to move the data all the way to the CPU cores. In this case, programmers would load their code onto the PIM cores and execute it there where the data can be fetched quickly, instead of executing it on the CPU cores. Sequence alignment algorithms are well-suited for such a system because there are usually a few computations performed for each data element accessed from the intermediate data structures (e.g. entries of the DP table in NW).

The UPMEM PIM architecture (Devaux 2019) is the first publicly available general-purpose programmable (<https://sdk.upmem.com>) processing-in-DRAM architecture. An UPMEM system consists of a set of UPMEM DIMM modules plugged alongside main

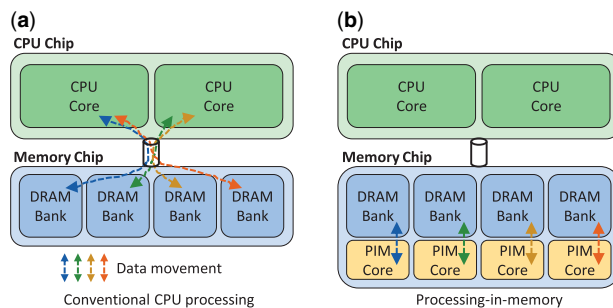


Figure 1 Comparison of conventional CPU processing and PIM.

memory (standard DDR4 DIMMs) and acting as parallel co-processors to the host CPU. An UPMEM module is a DDR4-2400 DIMM with 16 PIM-enabled chips, where each chip consists of eight general-purpose processing PIM cores called *DRAM Processing Units (DPUs)*. Each DPU is coupled with a 64-MB DRAM bank called *main RAM (MRAM)*. A current UPMEM system supports up to 20 UPMEM modules, which is equivalent to 2560 DPUs and 160 GB of memory. A DPU is a 32-bit RISC processor with a proprietary Instruction Set Architecture (ISA), which can potentially run at 500 MHz. Each DPU has 24 hardware threads that share a 24-KB instruction memory (IRAM) and a 64-KB scratchpad memory which is also called a working RAM (WRAM). The threads also share the 64-MB MRAM bank coupled with the DPU. DPUs cannot communicate with other DPUs or access data outside their own MRAM bank. The host CPU transfers data between the main memory and the MRAM banks and coordinates communication and synchronization across DPUs if needed.

### 2.2 Overall workflow

Figure 2 illustrates our framework’s overall workflow. In Step (1), we read the sequence pairs from an input file on disk and store them to the main memory. In Step (2), we transfer the sequence pairs from the main memory to the UPMEM DIMMs, distributing them evenly across the MRAM banks of the different DPUs. We use parallel transfers so that the sequence pairs are written to multiple MRAM banks simultaneously, thereby optimizing the transfer latency.

Next, we launch the DPU kernels and have each thread in each DPU work independently to align a set of sequence pairs. This parallelization scheme avoids inter-DPU and inter-thread synchronization, which can be expensive in the UPMEM system (Gómez-Luna et al. 2021a). In Step (3), each DPU thread performs a DMA transfer to fetch a sequence pair from MRAM and store it in WRAM. In Step (4), the DPU thread aligns the sequence pair and extracts the alignment operations using traceback. Our framework supports five different alignment algorithms: NW, SWG, GenASM, WFA, and WFA adaptive. In Step (5), the DPU thread performs a DMA transfer to write the alignment score and operations to MRAM. The thread then moves on to process the next sequence pair, repeating Steps (3)–(5) until all sequence pairs have been processed.

Finally, once all DPUs finish execution, we retrieve and save the alignment results. In Step (6), we transfer the alignment score and operations of each sequence pair from the UPMEM DIMMs to main memory using parallel transfers. In Step (7), we load the results from main memory and write them to an output file on disk.

### 2.3 Data management

One important aspect of implementing alignment algorithms on the UPMEM PIM architecture efficiently is data management. Recall from Section 2.1 that an UPMEM DPU has access to two memory

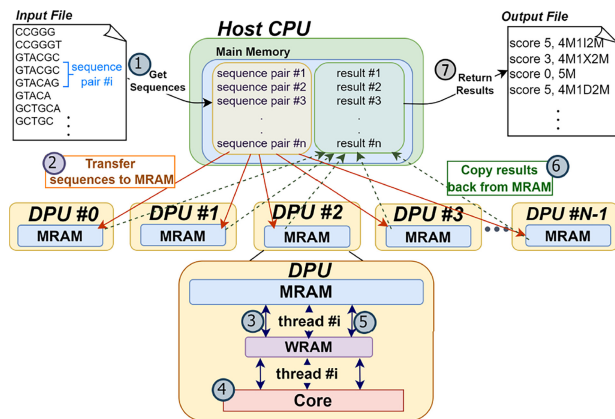


Figure 2 Overall workflow of AIM.

spaces for data: a 64-KB WRAM and a 64-MB MRAM. WRAM is fast and is accessed via loads and stores, whereas MRAM is slow and is accessed via DMA transfers to and from WRAM. Our framework always places the full sequence pair to be aligned and the full alignment result in WRAM for fast access because these data items are small. However, the intermediate data structures used by the alignment algorithms are relatively large. For this reason, it may not be possible to fit the entire intermediate data structure for each DPU thread in WRAM while supporting a large enough number of DPU threads to efficiently utilize the DPU pipeline. In this case, the constrained WRAM capacity can act as a limiting factor to parallelism.

To tackle this trade-off, we provide two alternative implementations of each alignment algorithm. This first implementation, illustrated in Fig. 3 (top), places the entire intermediate data structure for each DPU thread in WRAM, thereby prioritizing fast load/store access to the intermediate data structures. However, as the aligned sequences get larger, the WRAM capacity begins to constrain the number of DPU threads that can be launched, which causes the DPU pipeline to be underutilized. The second implementation, illustrated in Fig. 3 (bottom), places the intermediate data structure for each DPU thread in MRAM and transfers the parts of the data structure that need to be accessed to WRAM on-demand. With the WRAM capacity less of an issue, this approach enables using a larger number of DPU threads to better utilize the DPU pipeline. However, it incurs longer access latency to the intermediate data structures. The user of our framework can easily select which implementation they would like to use, and our framework automatically adapts the number of DPU threads launched depending on the implementation selected as well as the choice of alignment algorithm, read length, and error rate. We evaluate the trade-off between these two implementations in Section 3.5.

Another important aspect of data management is performing dynamic memory allocation efficiently. NW, SWG, and GenASM have fixed-size data structures that are allocated in WRAM or MRAM up-front. However, WFA and WFA adaptive rely on dynamic memory allocation which is performed frequently as the algorithms run. Dynamic memory allocation is needed for allocating the wavefront components, which vary in size at run time depending on the read length and similarity. The UPMEM SDK has an incremental dynamic memory allocator which allocates memory incrementally from beginning to end and then frees it all at once. However, it is not suitable for our purpose because it is shared by all threads so it

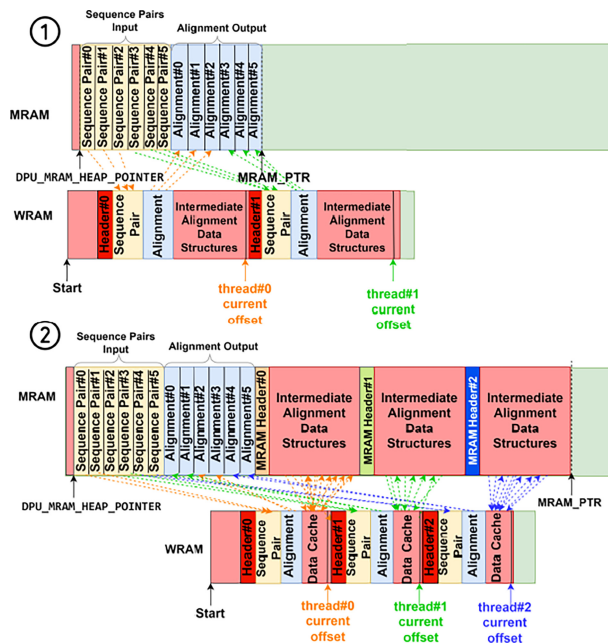


Figure 3 Example of using WRAM only (top) or using both WRAM and MRAM (bottom) for intermediate alignment data structures.



requires synchronization across threads to perform allocation, which degrades performance. To overcome this issue, we provide our own per-thread custom memory allocator to perform low-overhead dynamic memory allocation in the WFA and WFA-adaptive algorithms. The allocator also ensures that allocations are properly aligned such that they can be used in DMA transfers between WRAM and MRAM.

## 2.4 Supported alignment algorithms

### 2.4.1 Needleman–Wunsch

NW (Needleman and Wunsch 1970) computes the alignment of sequences using a DP table. In the implementation that only uses WRAM for intermediate alignment data structures, we have each DPU thread allocate its entire DP table in WRAM, fill it, perform the traceback, and send the alignment result to MRAM. The size of the DP table is  $m \cdot n$  where  $m$  and  $n$  are the lengths of the two aligned sequences. The data type of the DP-cells is `int16`. Therefore, the WRAM memory consumption per thread consists of the DP table ( $m \cdot n \cdot \text{sizeof}(\text{int16})$ ), the sequence pair, and the traceback operations. In this case, the 64-KB WRAM capacity limits the maximum read length to 175 bp, where one DPU thread is executed consuming 61-KB of WRAM. On the other hand, in the implementation that uses MRAM for intermediate alignment data structures, we store the DP table of each DPU thread in MRAM and use WRAM to store the neighboring DP-table cells of the current cell being computed. In this case, the 64-MB MRAM capacity limits the maximum read length to around 4 kb. NW uses the linear gap model to compute the alignment score. In our evaluation, we set the scoring parameters to  $a = 0$  (match cost),  $x = 3$  (mismatch cost), and  $e = 4$  (deletion/insertion cost).

### 2.4.2 Smith–Waterman–Gotoh

SWG (Gotoh 1982) resembles NW but uses an affine gap model which treats opening new gaps and extending existing gaps differently. To do so, SWG uses three DP tables—Matching (M), Insertion (I), and Deletion (D)—each of size  $m \cdot n$ . In the implementation that only uses WRAM for intermediate alignment data structures, we have each DPU thread store all three DP tables in WRAM. Since the memory usage of SWG is higher than that of NW, the maximum read length that can be used in this case is only 100 bp. On the other hand, in the implementation that uses MRAM for intermediate alignment data structures, we store the three DP tables of each DPU thread in MRAM and use WRAM to store the neighboring DP-table cells of the current cell being computed. In this case, the maximum read length that can be used is around 2.5 kb. We physically store the three tables as a single table where each cell has three consecutive values. By doing so, we can transfer cells from all three tables with the same DMA transfer, thereby amortizing the cost of transferring data from MRAM over fewer DMA transfers. In our evaluation, we set the scoring parameters to  $a = 0$  (match cost),  $x = 3$  (mismatch cost),  $o = 4$  (deletion/insertion opening cost), and  $e = 1$  (deletion/insertion extension cost).

### 2.4.3 GenASM

GenASM (Cali et al. 2020) is a recently proposed alignment algorithm that modifies and adds a traceback method to the bitap algorithm (Baeza-Yates and Gonnet 1992, Wu and Manber 1992). It uses the affine gap model and takes as an input the maximum number of edit distances ( $k$ ) allowed while computing the alignment. In the implementation that only uses WRAM for intermediate alignment data structures, we have each DPU thread use WRAM to store the pattern bit-mask for each character in the alphabet, two status bit-vectors to hold the partial alignment between subsequences of the sequences in the pair, and four intermediate bit-vectors for each edit case (matching, substitution, deletion, and insertion). On the other hand, in the implementation that uses MRAM for intermediate alignment data structures, we store the pattern bit-mask and the two status bit-vectors in MRAM and transfer parts of them to WRAM as needed. However, the four intermediate bit vectors are

still allocated in WRAM since they are small in size. In our evaluation, we set  $k$  according to the read length and error rate used.

### 2.4.4 Wavefront algorithm

WFA (Marco-Sola et al. 2020a,b) is the state-of-the-art affine gap alignment algorithm that computes exact pairwise alignments efficiently using wavefronts in the DP table. Each wavefront represents an alignment score, and the algorithm finds successive wavefronts (i.e. computes increasing-score partial alignments) until reaching the optimal alignment. Hence, the complexity of WFA is  $O(n \cdot s)$ , where  $s$  is the alignment score. As the alignment score  $s$  increases, WFA takes longer to execute and consumes more memory because it spans more diagonals. For this reason, WFA adaptive (Marco-Sola et al. 2020a,b), a heuristic variant of WFA, reduces the number of the spanned diagonals by eliminating outer diagonals that are unlikely to lead to an optimal alignment. Note that both WFA and WFA adaptive have linear space complexity, which makes them representative of other linear space sequence alignment algorithms (Myers and Miller 1988, Durbin et al. 1998) that could also benefit from our framework.

In our framework, we provide implementations of both WFA and WFA adaptive. In the implementations that only use WRAM for intermediate alignment data structures, we have each DPU thread use WRAM to store all the wavefront components. On the other hand, in the implementations that use MRAM for intermediate alignment data structures, we store all the wavefront components in MRAM and keep the addresses of the components in WRAM so they can be found when needed. To compute a new wavefront component  $WF_s$ , a DPU thread transfers from MRAM to WRAM only the components it needs. After computing  $WF_s$ , the DPU thread transfers the result from WRAM to MRAM. In our evaluation, we set the scoring parameters to  $a = 0$ ,  $x = 3$ ,  $o = 4$ , and  $e = 1$ .

## 3 Evaluation

### 3.1 Experimental setup

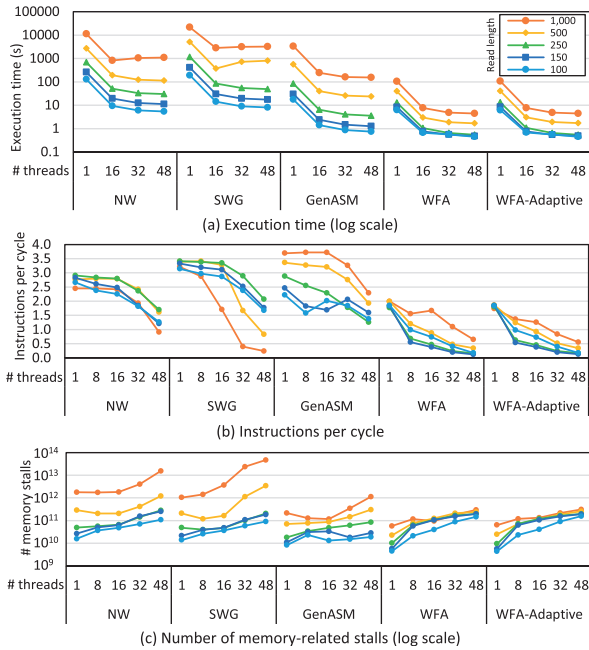
We compare the performance of our proposed framework to a multi-threaded CPU baseline that uses OpenMP to align multiple sequence pairs in parallel. The baseline CPU implementations of NW, SWG, WFA, and WFA adaptive are taken from the original WFA repository (Marco-Sola et al. 2020a,b), and the baseline CPU implementation of GenASM is taken from the GenASM repository (Cali et al. 2020).

We evaluate our PIM implementations on a UPMEM system with 2560 DPUs (20 UPMEM-DIMMs) running at 425 MHz. We evaluate the CPU implementations on three different *server-grade* CPU systems: (i) a dual socket Intel Xeon Silver 4215 CPU with 32 threads (16 cores, 8 cores/socket), 2.50 GHz frequency, 22 MB of L3 cache (11 MB/socket), and 256 GB of main memory, (ii) a dual socket Intel Xeon Gold 5120 CPU with 56 threads (28 cores, 14 cores/socket), 2.20 GHz frequency, 38 MB of L3 cache (19 MB/socket), and 64 GB of main memory, and (iii) a dual socket Intel Xeon E5-2697 v2 CPU with 48 threads (24 cores, 12 cores/socket), 2.70 GHz frequency, 60 MB of L3 cache (30 MB/socket), and 32 GB of main memory. We use execution time as the basis for comparison, which includes the wall clock time for performing data transfer, alignment, backtrace, and CIGAR string generation.

We use real and synthetic datasets to evaluate our proposed framework, as shown in Table 1. The real datasets are short read-reference pairs generated using minimap2 (Li 2018) by mapping the datasets (<https://www.ebi.ac.uk/ena/browser/view>) mentioned in Table 1 to the human reference genome GRCh37 (church2011modernizing). The synthetic datasets are long sequence pairs simulated using the synthetic data generator provided in the WFA repository (Marco-Sola et al. 2020a,b).

**Table 1.** Datasets used for the evaluation.

| Read lengths            | Edit distances (%) | Description                        |
|-------------------------|--------------------|------------------------------------|
| 100                     | 0–5                | Real, Accession # ERR240727        |
| 150                     | 0–5                | Real, Accession # SRR826460        |
| 250                     | 0–5                | Real, Accession # SRR826471        |
| 500, 1000, 5000, 10 000 | 0–5                | Synthetic (Marco-Sola et al. 2020) |

**Figure 4** Evaluation of CPU implementations on the best performing CPU system (Intel Xeon E5) while aligning five million sequence pairs.

### 3.2 CPU performance

Figure 4a shows how the execution times of the CPU implementations scale with the number of CPU threads while aligning five million sequence pairs using different alignment algorithms and read lengths, and an edit distance of 1%. The results are reported for the Xeon E5 CPU system, which is the best performing CPU system as we show in Section 3.3. Our key observation is that the CPU implementations face limited performance improvement as the number of CPU threads grows. To investigate the cause of this limited performance improvement, we profile the application using `perf`. Figure 4b shows that as the number of CPU threads increases, the instructions executed per cycle by each thread decreases. Hence, the CPU is increasingly underutilized when more threads are added. To further understand what is causing the CPU to be underutilized, Fig. 4c shows that as the number of CPU threads increases, the cycles spent by the CPU stalling while waiting for memory requests increases. Hence, the limited performance improvement is caused by the inability of the memory to serve memory requests quickly enough. These results demonstrate the importance of using PIM to overcome the memory bandwidth bottleneck faced by sequence alignment applications.

### 3.3 PIM performance versus CPU performance

Figure 5 shows the execution time of our framework aligning five million sequence pairs using different alignment algorithms, read lengths, and edit distances. We report the execution time both with and without the data transfer time between main memory and the UPMEM DIMMs. Based on these results, we make four key observations.

The first observation is that among the three CPU baselines, the best performing baseline in the majority of cases is the Xeon E5

CPU system. Despite not having the largest number of threads, this baseline has the largest L3 cache. This result demonstrates the memory boundedness of the sequence alignment problem, where having a larger L3 cache is favored over more having more threads.

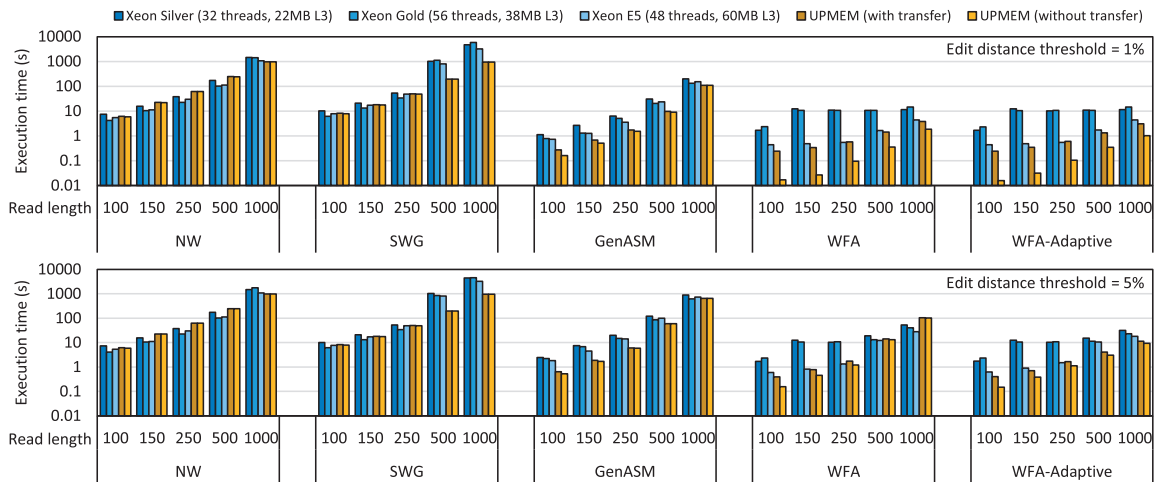
The second observation is that our framework running on the UPMEM system outperforms the CPU baselines in the majority of cases, even when data transfer time is included. The speedup achieved over the best CPU baseline is up to  $4.06\times$  in the case of SWG. For the state-of-the-art algorithms, WFA and WFA adaptive, the speedups achieved over the best CPU baseline are up to  $1.83\times$  and  $2.56\times$ , respectively. These results demonstrate the effectiveness of PIM at accelerating memory-bound sequence alignment workloads.

Here, we would like to reiterate that our CPU baselines are powerful server-grade dual-socket systems running at full scale. CPU hardware and software have been developed and optimized for decades by large teams of engineers, whereas the UPMEM PIM system has been developed over a few years by a small team. We expect that the relative advantage of PIM systems will be more pronounced as PIM hardware and software matures. We also note that in our current system, the DPUs are running at 425 MHz; however, they are expected to run at 500 MHz in future systems which would further improve performance.

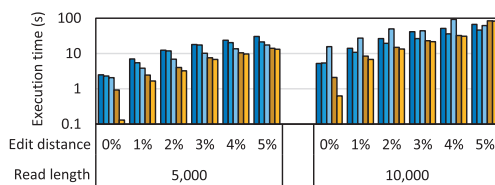
The third observation is that, when the data transfer time is *not* included, our framework achieves a speedup over the best CPU baseline of up to  $28.14\times$  in the case of WFA adaptive ( $25.93\times$  for WFA). The results without transfer time are important for two reasons. The first reason is that in the current UPMEM system, the UPMEM DIMMs cannot be used as regular main memory. For this reason, the CPU must read the data from disk to main memory then transfer it from main memory to the UPMEM DIMMs. However, in future systems where a PIM module could also be used as main memory, the CPU could potentially read data from disk to the PIM module directly and then perform the alignment in the PIM module without the additional transfer. The second reason is that in the current UPMEM system, one cannot overlap writing to MRAM by the CPU and execution by the DPUs. However, in future systems where CPU access to a PIM module may be overlapped with execution in the PIM cores, such overlapping can hide some of the latency of writing to the PIM module. Therefore, in light of the two aforementioned reasons, the results without the transfer time demonstrate the potential that future PIM systems have for accelerating memory-bound sequence alignment workloads.

The fourth observation is that our framework does not outperform the best CPU baseline for NW and SWG at small read lengths. NW and SWG have more regular memory access patterns than the other alignment algorithms, and when the read lengths are small, the intermediate data structures fit into the CPU cache. The combination of these two factors makes the memory bandwidth bottleneck less pronounced, which diminishes the advantage of PIM. However, for other algorithms where the memory access pattern is less regular, and for larger read lengths where the sizes of the intermediate data structures outgrow the size of the CPU cache, the computation becomes more memory bound causing our framework to achieve large speedups over the CPU baseline.

To further study the scalability of our framework, we evaluate the state-of-the-art algorithm, WFA adaptive, on large sequence lengths of 5 and 10 kb. NW and SWG have difficulty scaling to such large read lengths because their quadratic space complexity causes them to exceed the MRAM capacity even for a single alignment. However, WFA and WFA adaptive are not limited by MRAM capacity due to their linear space complexity. The execution time of



**Figure 5** Execution time of our framework while aligning five million sequence pairs using different alignment algorithms, read lengths, and edit distances compared with three CPU baselines.



**Figure 6** Execution time of our framework aligning one million sequence pairs using WFA adaptive for large read lengths compared with three CPU baselines (same legend as Fig. 5).

WFA adaptive for large read lengths while aligning one million sequence pairs is shown in Fig. 6.

We observe that our framework executing on the UPMEM system continues to outperform all CPU baselines. The only exception is for read length 10 000 with an edit distance threshold of 5%. In this case, the WRAM capacity is only sufficient to support a single DPU thread per DPU which underutilizes the DPU pipelines. Although the algorithm is not limited by the 64-MB MRAM for storing all the wavefront components, it is limited by the 64-KB WRAM for storing the select wavefront components needed for computing a new wavefront component. Hence, our current framework executing on the current hardware cannot scale to read lengths far beyond 10 000 for WFA and WFA adaptive due to WRAM capacity constraints. In some bioinformatics applications (such as minimap2), sequence alignment is typically performed only between every two seed chains to avoid long execution time and peak memory that could result from performing sequence alignment on the complete read sequence. Thus, we believe that our framework is useful for a wide range of applications despite the sequence length limitation. However, if sequence lengths far beyond 10 000 are a concern, the current limitation can be mitigated by streaming partial wavefront components from MRAM to WRAM in order to reduce the WRAM footprint and support aligning more sequence pairs of larger length simultaneously, or by using multiple DPU threads to align a single sequence pair. These improvements are the subject of our future work. We also expect the supportable read length and the performance for large read lengths to improve in future systems as the hardware and software mature, as the clock frequency of the DPUs increases, and as future systems may have larger WRAM capacity.

### 3.4 PIM performance versus GPU performance

Table 2 compares the throughput of our PIM implementations of WFA adaptive (the fastest algorithm) to a recent GPU implementation of WFA adaptive by Aguado-Puig et al. (2022) using the results reported in that work. It is clear that our PIM implementation outperforms the GPU implementation in the majority of cases. This result shows that PIM is a

promising technology for accelerating sequence alignment, even when compared to mature and widely used accelerators such as GPUs.

### 3.5 Using WRAM only or WRAM and MRAM

Recall from Section 2.3 that our framework provides two implementations of each algorithm: one that only uses WRAM for intermediate alignment data structures, and another that uses MRAM for these data structures and transfers data currently being accessed from these data structures to WRAM as needed. Figure 7 compares the execution time and scalability of these two implementations for each algorithm and read length with edit distance 1%. Based on these results, we make three key observations.

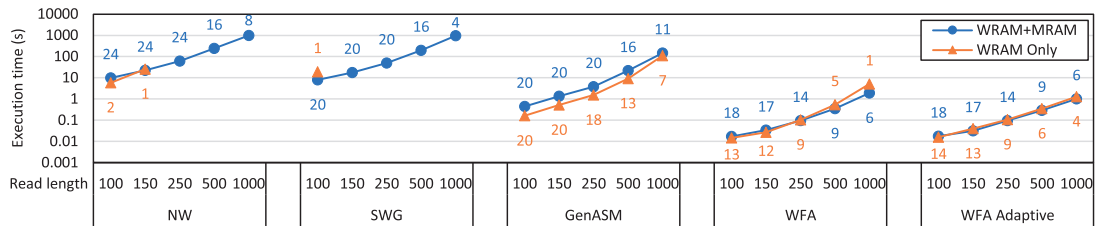
The first observation is that for the algorithms that use a large amount of memory for the intermediate alignment data structures (i.e. NW and SWG), the implementations that use both WRAM and MRAM scale better with the read length than those that only use WRAM. In the case of NW for small read lengths, the implementation that uses WRAM only is up to  $1.70\times$  faster. However, for the remaining cases, this implementation is slower or cannot even execute. The reason is that the implementations that use WRAM only can only support a small number of DPU threads due to the constrained WRAM capacity, which prevents them from utilizing the DPU pipeline well. In contrast, the implementations that use both WRAM and MRAM can support a larger number of DPU threads, causing them to perform better for long read lengths despite incurring higher memory access latency.

The second observation is that for the algorithm that uses a small amount of memory for the intermediate alignment data structures (i.e. GenASM), the implementation that only uses WRAM is faster (up to  $2.76\times$ ). The reason is that the WRAM only implementation can support a large enough number of DPU threads to utilize the DPU pipeline well. We note, however, that for larger edit distances that require more memory, the implementation that uses WRAM and MRAM becomes more favorable for large reads. For example, for read length 1000 and edit distance 5% (not shown in Fig. 7), the implementation that uses WRAM and MRAM together is  $1.91\times$  faster.

The third observation is that for the implementations that use a moderate amount of memory for the intermediate alignment data structures (i.e. WFA and WFA adaptive), the implementations that use WRAM only are faster for shorter reads (up to  $1.17\times$  for WFA and  $1.12\times$  for WFA adaptive). On the other hand, the implementations that use both WRAM and MRAM are faster for longer reads (up to  $2.70\times$  for WFA and  $1.25\times$  for WFA adaptive). The reason is that the number of DPU threads that can be used decreases as the read length gets larger, which favors the implementation that can use more threads over the one with lower access latency. Note that the difference between the two implementations grows as the edit distance grows because the memory consumption of WFA and WFA

**Table 2.** Comparison with WFA-GPU (Aguado-Puig et al. 2022).

| Sequence length | Edit distance (%) | Throughput (alignments per second) |                       | Throughput improvement |
|-----------------|-------------------|------------------------------------|-----------------------|------------------------|
|                 |                   | WFA-GPU                            | UPMEM (with transfer) |                        |
| 150             | 2                 | 9.09M                              | 12.97M                | 1.42×                  |
|                 | 5                 | 5.56M                              | 7.03M                 | 1.27×                  |
| 1000            | 2                 | 1.43M                              | 1.10M                 | 0.77×                  |
|                 | 5                 | 370K                               | 434K                  | 1.17×                  |
| 10 000          | 2                 | 25.0K                              | 66.9K                 | 2.68×                  |
|                 | 5                 | 5.56K                              | 11.81K                | 2.12×                  |

**Figure 7** Execution time of our framework aligning five million sequence pairs with edit distance 1% using WRAM Only or WRAM and MRAM (labels indicate number of threads per DPU).

adaptive is highly sensitive to the edit distance. For example, for edit distance 5% (not shown in the plot), the implementation that uses WRAM and MRAM together is up to  $6.15\times$  faster.

The three observations presented in this section demonstrate the importance of our framework supporting both implementations of each algorithm, those that use WRAM only and those that use WRAM and MRAM for intermediate alignment data structures. In either case, our framework can automatically identify the maximum number of threads that can be used to alleviate this burden from the user.

## 4 Conclusion

We present a framework for high-throughput pairwise sequence alignment that overcomes the memory bandwidth bottleneck by using real PIM systems. Our framework targets UPMEM, the first publicly available general-purpose programmable PIM architecture. It supports multiple alignment algorithms and includes two implementations of each algorithm that manage the UPMEM memory hierarchy differently and are suitable for different read lengths. Our evaluation shows that our framework executing on an UPMEM PIM system substantially outperforms parallel CPU implementations executing at full-scale on dual-socket server-grade CPU systems. Our results demonstrate that PIM systems provide a promising alternative for accelerating sequence alignment. We expect even larger improvements from future incarnations of PIM systems.

## Acknowledgments

We thank Fabrice Devaux and Remy Cimadomo for their valuable support, including insightful feedback and access to UPMEM hardware.

## Conflict of interest

None declared.

## Funding

This work was supported by the University Research Board of the American University of Beirut [URB-AUB-104107-26306]. We also acknowledge the generous gifts provided by the SAFARI Research Group's industrial partners, including ASML, Facebook, Google, Huawei, Intel, Microsoft, VMware, and Xilinx, as well as the support from the Semiconductor Research Corporation,

the ETH Future Computing Laboratory, and the European Union's Horizon programme for research and innovation under grant agreement No 101047160, project BioPIM (Processing-in-memory architectures and programming libraries for bioinformatics algorithms).

## References

- Aguado-Puig, Q., Marco-Sola, S., Moure, J. C., et al. (2022). WFA-GPU: Gap-affine pairwise alignment using GPUs. *bioRxiv*.
- Ahmed, N., Qiu, T. D., Bertels, K., et al. (2020). GPU acceleration of Darwin read overlap for de novo assembly of long DNA reads. *BMC bioinformatics*, 21(13).
- Alser, M., Hassan, H., Xin, H., et al. (2017a). GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping. *Bioinformatics*, 33(21), 3355–3363.
- Alser, M., Mutlu, O., and Alkan, C. (2017b). MAGNET: Understanding and improving the accuracy of genome pre-alignment filtering. *Transactions on Internet Research*, 13(2), 33–42.
- Alser, M., Hassan, H., Kumar, A., et al. (2019). Shouji: a fast and efficient pre-alignment filter for sequence alignment. *Bioinformatics*, 35(21), 4255–4263.
- Alser, M., Bingöl, Z., Cali, D. S., et al. (2020a). Accelerating genome analysis: A primer on an ongoing journey. *IEEE Micro*, 40(5), 65–75.
- Alser, M., Shahroodi, T., Gómez-Luna, J., et al. (2020b). SneakySnake: a fast and accurate universal genome pre-alignment filter for CPUs, GPUs and FPGAs. *Bioinformatics*, 36(22-23), 5282–5290.
- Alser, M., Rotman, J., Taraszka, K., et al. (2020c). *Technology dictates algorithms: Recent developments in read alignment*. arXiv preprint arXiv:2003.00110.
- Alser, M., Lindegger, J., Firtina, C., et al. (2022). *Going from molecules to genomic variations to scientific discovery: Intelligent algorithms and architectures for intelligent genome analysis*. arXiv preprint arXiv:2205.07957.
- Ankit, A., El Hajj, I., Chalamalasetti, S. R., et al. (2019). PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference. In: *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 715–731. New York, NY, USA: Association for Computing Machinery.
- Ankit, A., El Hajj, I., Chalamalasetti, S. R., et al. (2020). PANTHER: A programmable architecture for neural network training harnessing energy-efficient reram. *IEEE Transactions on Computers*, PP.
- Arlazarov, V. L., Dinitz, Y. A., Kronrod, M., et al. (1970). On economical construction of the transitive closure of an oriented graph. In *Doklady Akademii Nauk*, volume 194, pp. 487–488. Russian Academy of Sciences.
- Backurs, A. and Indyk, P. (2015). Edit distance cannot be computed in strongly subquadratic time (unless seth is false). In: *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 51–58. New York, NY, USA: Association for Computing Machinery.



- Baeza-Yates, R. and Gonnet, G. H. (1992). A new approach to text searching. *Commun. ACM*, 35(10), 74–82.
- Banerjee, S. S., El-Hadedy, M., Lim, J. B., et al. (2018). ASAP: accelerated short-read alignment on programmable hardware. *IEEE Transactions on Computers*, 68(3), 331–346.
- Cali, D. S., Kalsi, G. S., Bingöl, Z., et al. (2020). GenASM: A high-performance, low-power approximate string matching acceleration framework for genome sequence analysis. In: *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 951–966. IEEE.
- Cali, D. S., Kanellopoulos, K., Lindegger, J., et al. (2022). Segram: A universal hardware accelerator for genomic sequence-to-graph and sequence-to-sequence mapping. In: *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp. 638–655. New York, NY, USA: Association for Computing Machinery.
- Church, D. M., Schneider, V. A., Graves, T., et al. (2011). Modernizing reference genome assemblies. *PLoS biology*, 9(7), e1001091.
- Daily, J. (2016). Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC bioinformatics*, 17(1), 1–11.
- Devaux, F. (2019). The true processing in memory accelerator. In: *2019 IEEE Hot Chips 31 Symposium (HCS)*, Cupertino, CA, USA, pp. 1–24. IEEE Computer Society.
- Diab, S., Nassereldine, A., Alser, M., et al. (2022). High-throughput pairwise alignment with the wavefront algorithm using processing-in-memory. arXiv:2204.02085.
- Durbin, R., Eddy, S. R., Krogh, A., et al. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press.
- Ferreira, J. D., Falcao, G., Gómez-Luna, J., et al. (2022). pluto: Enabling massively parallel computation in dram via lookup tables. In: *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Chicago, IL, USA, pp. 900–919. IEEE.
- Fujiki, D., Subramaniyan, A., Zhang, T., et al. (2018). Genax: A genome sequencing accelerator. In: *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, Los Angeles, CA, USA, pp. 69–82. IEEE.
- Giannoula, C., Fernandez, I., Luna, J. G., et al. (2022). Sparsep: Towards efficient sparse matrix vector multiplication on real processing-in-memory architectures. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1), 1–49.
- Gómez-Luna, J., El Hajj, I., Fernandez, I., et al. (2021a). Benchmarking a new paradigm: An experimental analysis of a real processing-in-memory architecture. arXiv preprint arXiv:2105.03814.
- Gómez-Luna, J., El Hajj, I., Fernandez, I., et al. (2021b). Benchmarking memory-centric computing systems: Analysis of real processing-in-memory hardware. In: *12th International Green and Sustainable Computing Conference (IGSC)*, Pullman, WA, USA, 2021, pp. 1–7. IEEE.
- Gómez-Luna, J., El Hajj, I., Fernandez, I., et al. (2022). Benchmarking a new paradigm: Experimental analysis and characterization of a real processing-in-memory system. *IEEE Access*, 10, 52565–52608.
- Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3), 705–708.
- Gupta, S., Imani, M., Khaleghi, B., et al. (2019). RAPID: A ReRAM processing in-memory architecture for DNA sequence alignment. In: *IEEE/ACM International Symposium on Low Power Electronics and Design, Lausanne, Switzerland*, pp. 1–6. IEEE.
- Haghi, A., Marco-Sola, S., Alvarez, L., et al. (2021). An FPGA accelerator of the wavefront algorithm for genomics pairwise alignment. In *2021 31st International Conference on Field-Programmable Logic and Applications (FPL)*, pages 151–159.
- Hajinazar, N., Oliveira, G. F., Gregorio, S., et al. (2021a). SIMDRAM: a framework for bit-serial SIMD processing using DRAM. In: *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 329–345. New York, NY, USA: Association for Computing Machinery.
- Hajinazar, N., Oliveira, G. F., Gregorio, S., et al. (2021b). SIMDRAM: An end-to-end framework for bit-serial SIMD computing in DRAM. arXiv preprint arXiv:2105.12839.
- Huang, S., Ankit, A., Silveira, P., et al. (2021). Mixed precision quantization for reram-based dnn inference accelerators. In: *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Tokyo, Japan, pp. 372–377. IEEE.
- Hwu, W.-m., El Hajj, I., De Gonzalo, S. G., et al. (2017). Rebooting the data access hierarchy of computing systems. In: *2017 IEEE International Conference on Rebooting Computing (ICRC)*, Washington, DC, pp. 1–4. IEEE.
- Kalikar, S., Jain, C., Vasimuddin, M., et al. (2022). Accelerating minimap2 for long-read sequencing applications on modern CPUs. *Nature Computational Science*, 2(2), 78–83.
- Kaplan, R., Yavits, L., and Ginosar, R. (2020). BioSEAL: In-memory biological sequence alignment accelerator for large-scale genomic data. In: *13th ACM International Systems and Storage Conference*, pp. 36–48. New York, NY, USA: Association for Computing Machinery.
- Kim, J. S., Cali, D. S., Xin, H., et al. (2018). GRIM-Filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies. *BMC genomics*, 19(2), 23–40.
- Lavenier, D., Deltel, C., Furodet, D., et al. (2016a). BLAST on UPMEM. Ph.D. thesis, INRIA Rennes-Bretagne Atlantique.
- Lavenier, D., Roy, J.-F., and Furodet, D. (2016b). DNA mapping using processor-in-memory architecture. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, China, pp. 1429–1435. IEEE.
- Lavenier, D., Deltel, C., Furodet, D., et al. (2016c). MAPPING on UPMEM. Ph.D. thesis, INRIA.
- Lavenier, D., Jodin, R., and Cimadomo, R. (2020). *Variant calling parallelization on processor-in-memory architecture*. bioRxiv.
- Li, H. (2018a). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100.
- Li, H. (2018b). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100.
- Mansouri Ghiasi, N., Park, J., Mustafa, H., et al. (2022). Genstore: a high-performance in-storage processing system for genome sequence analysis. In: *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 635–654. New York, NY, USA: Association for Computing Machinery.
- Marco-Sola, S., Moure López, J. C., Moreto Planas, M., et al. (2020). Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*, (btaa777), 1–8.
- Marco-Sola, S., Eizenga, J. M., Guarracino, A., et al. (2022). Optimal gap-affine alignment in o(s) space. bioRxiv.
- Moore, G. E. (1998). Cramping more components onto integrated circuits. *Proceedings of the IEEE*, 86(1), 82–85.
- Mutlu, O., Ghose, S., Gómez-Luna, J., et al. (2019). Processing data where it makes sense: Enabling in-memory computation. *Microprocessors and Microsystems*, 67, 28–41.
- Mutlu, O., Ghose, S., Gómez-Luna, J., et al. (2020). *A modern primer on processing in memory*. arXiv preprint arXiv:2012.03112.
- Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *Bioinformatics*, 4(1), 11–17.
- Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1), 31–88.
- Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3), 443–453.
- Nider, J., Mustard, C., Zoltan, A., et al. (2021). A case study of processing-in-memory in off-the-shelf systems. In: *USENIX Annual Technical Conference*, pp. 117–130. Berkeley, CA, USA: USENIX Association.
- Rasmussen, K. R., Stoye, J., and Myers, E. W. (2006). Efficient q-gram filters for finding all  $\epsilon$ -matches over a given length. *Journal of Computational Biology*, 13(2), 296–308.
- Rizk, G. and Lavenier, D. (2010). GASSST: global alignment short sequence search tool. *Bioinformatics*, 26(20), 2534–2540.
- Šošić, M. and Šikić, M. (2017). Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9), 1394–1395.
- Turakhia, Y., Bejerano, G., and Dally, W. J. (2018). Darwin: A genomics co-processor provides up to 15,000 x acceleration on long read assembly. *ACM SIGPLAN Notices*, 53(2), 199–213.
- Vasimuddin, M., Misra, S., Li, H., et al. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Rio de Janeiro, Brazil, pp. 314–324. IEEE.
- Wu, S. and Manber, U. (1992). Fast text searching: allowing errors. *Communications of the ACM*, 35(10), 83–91.
- Xin, H., Lee, D., Hormozdiari, F., et al. (2013). Accelerating read mapping with fasthash. In: *BMC genomics*, Vol. 14, pp. 1–13. Springer.
- Xin, H., Greth, J., Emmons, J., et al. (2015). Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping. *Bioinformatics*, 31(10), 1553–1560.
- Zhang, Z., Schwartz, S., Wagner, L., et al. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational biology*, 7(1-2), 203–214.
- Zois, V., Gupta, D., Tsotras, V. J., et al. (2018). Massively parallel skyline computation for processing-in-memory architectures. In: *27th International Conference on Parallel Architectures and Compilation Techniques*, pp. 1–12. New York, NY, USA: Association for Computing Machinery.