Concurrent Data Structures for Near-Memory Computing

Zhiyu Liu
Computer Science Department
Brown University
zhiyu_liu@brown.edu

Irina Calciu
VMware Research Group
icalciu@vmware.com

Maurice Herlihy
Computer Science Department
Brown University
mph@cs.brown.edu

Onur Mutlu
Computer Science Department
ETH Zürich
onur.mutlu@inf.ethz.ch

ABSTRACT

The performance gap between memory and CPU has grown exponentially. To bridge this gap, hardware architects have proposed near-memory computing (also called processing-in-memory, or PIM), where a lightweight processor (called a PIM core) is located close to memory. Due to its proximity to memory, a memory access from a PIM core is much faster than that from a CPU core. New advances in 3D integration and die-stacked memory make PIM viable in the near future. Prior work has shown significant performance improvements by using PIM for embarrassingly parallel and data-intensive applications, as well as for pointer-chasing traversals in *sequential* data structures. However, current server machines have hundreds of cores, and algorithms for concurrent data structures exploit these cores to achieve high throughput and scalability, with significant benefits over sequential data structures. Thus, it is important to examine how PIM performs with respect to modern *concurrent* data structures and understand how concurrent data structures can be developed to take advantage of PIM.

This paper is the first to examine the design of *concurrent* data structures for PIM. We show two main results: (1) naive PIM data structures *cannot* outperform state-of-the-art concurrent data structures, such as pointer-chasing data structures and FIFO queues, (2) novel designs for PIM data structures, using techniques such as combining, partitioning and pipelining, can outperform traditional concurrent data structures, with a significantly simpler design.

KEYWORDS

concurrent data structures; parallel programs; processing-in-memory; near-memory computing

1 NEAR-MEMORY COMPUTING

The performance gap between memory and CPU has grown exponentially. Memory vendors have focused mainly on improving memory capacity and bandwidth, sometimes even at the cost of higher memory access latencies [11, 12, 14, 35, 37–39, 42, 43]. To provide higher bandwidth with lower access latencies, hardware architects have proposed near-memory computing (also called *processing-in-memory*, or PIM), where a lightweight processor (called a PIM core) is located close to memory. A memory access from a PIM core is much faster than that from a CPU core. Near-memory computing is an old idea that has been intensively studied in the past (e.g., [17, 20, 21, 32, 34, 44, 45, 51]), but so far has not yet materialized. However, new advances in 3D integration and die-stacked memory likely make near-memory computing viable in the near future. For example, one PIM design [1, 2, 9, 53] assumes that memory is organized in multiple vaults, each having an in-order PIM core to manage it. These PIM cores can communicate through message passing, but do not share memory, and cannot access each other's vaults.

This new technology promises to revolutionize the interaction between computation and data, as it enables memory to become an active component in managing the data. Therefore, it invites a fundamental rethinking of basic data structures and promotes a tighter dependency between algorithmic design and hardware characteristics.

Prior work has already shown significant performance improvements by using PIM for embarrassingly parallel and data-intensive applications [1, 3, 29, 53, 54], as well as for pointer-chasing traversals [23, 30] in *sequential* data structures. However, current server machines have hundreds of cores, and algorithms for concurrent data structures exploit these cores to achieve high throughput and scalability, with significant benefits over sequential data structures (e.g., [19, 27, 46, 52]). Unlike prior work, we focus on *concurrent* data structures for PIM and we show that naive PIM data structures *cannot* outperform state-of-the-art concurrent data structures. In particular, the lower-latency access to memory provided by PIM cannot compensate for the loss of parallelism in data structure manipulation. For example, we show that even if a PIM memory access is two times faster than a CPU memory access, a sequential PIM linked-list is still slower than a traditional concurrent linked-list accessed in parallel by only three CPU cores.

Therefore, to be competitive with traditional concurrent data structures, PIM data structures need new algorithms and new approaches to leverage parallelism. As the PIM technology approaches fruition, it is crucial to investigate how to best utilize it to exploit the lower latencies, while still leveraging the vast amount of previous research related to concurrent data structures.

In this paper, we provide answers to the following key questions: 1) How do we design and optimize data structures for PIM? 2) How

SPAA '17, July 24-26, 2017, Washington DC, USA

DOI: http://dx.doi.org/10.1145/3087556.3087582

do these optimized PIM data structures compare to traditional CPU-managed concurrent data structures? To answer these questions, even before the hardware becomes available, we develop a simplified model of the expected performance of PIM. Using this model, we investigate two classes of data structures.

First, we analyze *pointer chasing data structures* (Section 4), which have a high degree of inherent parallelism and low contention, but which incur significant overhead due to hard-to-predict memory access patterns. We propose using techniques such as combining and partitioning the data across vaults to reintroduce parallelism for these data structures.

Second, we explore *contended data structures* (Section 5), such as FIFO queues, which can leverage CPU caches to exploit their inherent high locality. As they exploit the fast on-chip caches well, FIFO queues might not seem to be a good fit for leveraging PIM's faster memory accesses. Nevertheless, these data structures exhibit a high degree of contention, which makes it difficult, even for the most advanced data structures, to obtain good performance when many threads access the data concurrently. We use pipelining of requests, which can be done very efficiently in PIM, to design a new FIFO queue suitable for PIM that can outperform state-of-the-art concurrent FIFO queues [25, 41].

The contributions of this paper are as follows:

- We propose a simple and intuitive model to analyze the performance of PIM data structures and of concurrent data structures. This model considers the number of atomic operations, the number of memory accesses and the number of accesses that can be served from the CPU cache.
- Using this model, we show that the lower-latency memory accesses provided by PIM are *not* sufficient for sequential PIM data structures to outperform efficient traditional concurrent data structures.
- We propose new designs for PIM data structures using techniques such as combining, partitioning and pipelining. Our evaluations show that these new PIM data structures can outperform traditional concurrent data structures, with a significantly simpler design.

The paper is organized as follows. In Section 2, we briefly describe our assumptions about the hardware architecture. In Section 3, we introduce a simplified performance model that we use throughout this paper to estimate the performance of our data structures using the hardware architecture described in Section 2. In Sections 4 and 5, we describe and analyze our PIM data structures and use our model to compare them to prior work. We also use current DRAM architectures to simulate the behavior of our data structures and evaluate them compared to state-of-the-art concurrent data structures. Finally, we present related work in Section 6 and conclude in Section 7.

2 HARDWARE ARCHITECTURE

In an example architecture utilizing PIM memory [1, 2, 9, 53], multiple CPUs are connected to the main memory, via a shared crossbar network, as illustrated in Figure 1. The main memory consists of two parts—one is a standard DRAM accessible by CPUs, and the other, called the *PIM memory*, is divided into multiple partitions, called *PIM vaults* or simply *vaults*. According to the *Hybrid Memory Cube (HMC)* specification 1.0 [15], each HMC consists of 16 or 32 vaults and has a total size of 2GB or 4GB (so each vault's size is roughly 100MB).[1] We assume the same specifications in our PIM model, although the size of the PIM memory and the number of its vaults can be bigger. Each CPU core also has access to a hierarchy of L1 and L2 caches backed by DRAM, and a last level cache shared among multiple cores.

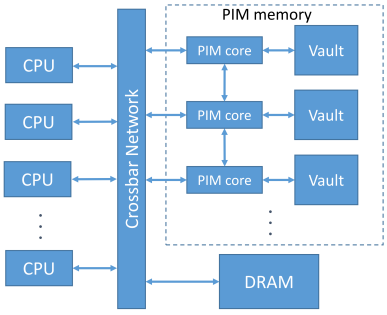

Figure 1: An example PIM architecture

Each vault has a *PIM core* directly attached to it. We say a vault is *local* to the PIM core attached to it, and vice versa. A PIM core is a lightweight CPU that may be slower than a full-fledged CPU with respect to computation speed [1]. A PIM core can be thought of as an in-order CPU with a small private L1 cache. A vault can be accessed only by its local PIM core.[2] Recent work proposes efficient cache coherence mechanisms between PIM cores and CPUs (e.g., [2, 9]), but this introduces additional complexity. We show that we can design efficient concurrent PIM data structures even if there is no coherence. Although a PIM core has lower performance than a state-of-the-art CPU core, it has fast access to its local vault.

A PIM core communicates with other PIM cores and CPUs via messages. Each PIM core, as well as each CPU, has buffers for storing incoming messages. A message is guaranteed to eventually arrive at the buffer of its receiver. Messages from the same sender to the same receiver are delivered in FIFO order: the message sent first arrives at the receiver first. However, messages from different senders or to different receivers can arrive in an arbitrary order.

We assume that a PIM core can only perform read and write operations to its local vault, while a CPU also supports more powerful atomic operations, such as *Compare-And-Swap (CAS)* and *Fetch-And-Add (F&A)*. Virtual memory can be realized efficiently if each PIM core maintains its own page table for the local vault [30].

3 PERFORMANCE MODEL

We propose the following simple performance model to compare our PIM-managed data structures with existing concurrent data structures. For read and write operations, we assume

$$\mathcal{L}_{cpu} = r_1 \mathcal{L}_{pim} = r_2 \mathcal{L}_{llc}$$

where $\mathcal{L}_{cpu}$ is the latency of a memory access by a CPU, $\mathcal{L}_{pim}$ is the latency of a local memory access by a PIM core, and $\mathcal{L}_{llc}$

[1] These small sizes are preliminary, and it is expected that each vault will become larger when PIM memory is commercialized.
[2] Alternatively, we could assume that a PIM core has direct access to the remote vaults, but such accesses are slower than those to the local vault.

is the latency of a last-level cache access by a CPU. Based on the latency numbers in prior work on PIM memory, in particular on the Hybrid Memory Cube [6, 15], and on the evaluation of operations in multiprocessor architectures [16], we may further assume

$$r_1 = r_2 = 3.$$

The latencies of operations may vary significantly on different machines. Our assumption that $r_1 = r_2 = 3$ is mainly to make the performance analysis later in the paper more concrete with actual latency numbers. In our performance model, we ignore the costs of accesses to other cache levels, such as L1 or L2, as they are negligible in the concurrent data structures we consider.

We assume that the latency of a CPU performing an atomic operation, such as a CAS or a F&A, to a cache line is

$$\mathcal{L}_{atomic} = r_3 \mathcal{L}_{cpu}$$

where $r_3 = 1$, even if the cache line is currently in the cache. This is because an atomic operation hitting in the cache is usually as costly as a memory access by a CPU [16]. When there are k atomic operations competing for a cache line concurrently, we assume that they are executed sequentially, that is, they complete in times $\mathcal{L}_{atomic}, 2\mathcal{L}_{atomic}, ..., k \cdot \mathcal{L}_{atomic}$, respectively.

We also assume that the size of a message sent by a PIM core or a CPU core is at most the size of a cache line. Given that a message transferred between a CPU and a PIM core goes through the crossbar network, we assume that the latency for a message to arrive at its receiver is

$$\mathcal{L}_{message} = \mathcal{L}_{cpu}$$

We make the conservative assumption that the latency of a message transferred between two PIM cores is also $\mathcal{L}_{message}$. Note that the message latency we consider here is the transfer time of a message through a message passing channel, that is, the elapsed time between the moment when a PIM or a CPU core finishes sending off the message and the moment when the message arrives at the buffer of its receiver. We ignore the time spent in other parts of a message passing procedure, such as in *preprocessing and constructing the message*, and in *actually sending the message*, as it is negligible compared to the time spent in the message transfer [6].

4 LOW-CONTENTION DATA STRUCTURES

In this section, we consider data structures with low contention. Pointer chasing data structures, such as linked-lists and skip-lists, fall in this category. These are data structures whose operations need to de-reference a non-constant sequence of pointers before completing. We assume these data structures support operations such as add(x), delete(x) and contains(x), which follow "next node" pointers until reaching the position of node x. When these data structures are too large to fit in the CPU caches and access uniformly random keys, they incur expensive memory accesses, which cannot be easily predicted, making the *pointer chasing* operations the dominating overhead of these data structures. Naturally, these data structures have provided early examples of the benefits of near-memory computing [23, 30], as the entire pointer chasing operation could be performed by a PIM core with fast memory access, and only the final result returned to the application.

However, these data structures have inherently low contention. Lock-free algorithms [19, 27, 46, 52] have shown that these data structures can scale to hundreds of cores under low contention [10]. Unfortunately, each vault in PIM memory has a single core. As a consequence, prior work has compared PIM data structures only with sequential data structures, *not* with carefully crafted concurrent data structures.

We analyze linked-lists and skip-lists, and show that the naive PIM data structure in each case cannot outperform the equivalent CPU-managed concurrent data structure even for a small number of cores. Next, we show how to use state-of-the art techniques from concurrent computing to optimize data structures for near-memory computing such that they outperform well-known concurrent data structures designed for multi-core CPUs.

4.1 Linked-lists

We first describe a naive PIM linked-list. The linked-list is stored in a vault, maintained by the local PIM core. Whenever a CPU[3] wants to perform an operation on the linked-list, it sends a request to the PIM core. The PIM core then retrieves the message, executes the operation, and sends the result back to the CPU. The PIM linked-list is sequential, as it can only be accessed by one PIM core.

Performing pointer chasing on sequential data structures using PIM cores is not a new idea. Prior work ([1, 23, 30]) has shown that pointer chasing can be done more efficiently by a PIM core for a sequential data structure. However, we are not aware of any prior comparison between the performance of PIM-managed data structures and *concurrent* data structures, for which CPUs can perform operations in parallel. In fact, our analytical and experimental results show that the naive PIM-managed linked-list is not competitive with a concurrent linked-list that uses fine-grained locks [24].

We use the *combining optimization* proposed by flat combining [25] to improve this data structure: a PIM core can execute all concurrent requests by CPU cores using *a single* traversal over the linked-list.

The role of the PIM core in our PIM-managed linked-list is very similar to that of the combiner in a concurrent linked-list implemented using *flat combining* [25], where, roughly speaking, threads compete for a "combiner lock" to become the combiner, and the combiner takes over all operation requests from other threads and executes them. Therefore, we consider the performance of the flat-combining linked-list as an indicator of the performance of our proposed PIM-managed linked-list.

Based on our performance model, we can calculate the approximate expected throughput (in operations per second) of each of the linked-lists mentioned above, when there are p CPUs making operation requests concurrently. We assume that a linked-list consists of nodes with integer keys in the range of $[1, N]$. Initially a linked-list has n nodes with keys generated independently and uniformly at random from $[1, N]$. The keys of the operation requests are generated the same way. To simplify the analysis, we assume that the size of the linked-list does not fluctuate much. This is achieved when the number of add() requests is similar to the number of delete() requests. We assume that a CPU makes a new operation request

[3] We use the term CPU to refer to a CPU core, as opposed to a PIM core.

immediately after its previous one completes. Assuming that $n \gg p$ and $N \gg p$, the approximate expected throughput (per second) of each of the concurrent linked-lists is presented in Table 1, where $\mathcal{S}_p = \sum_{i=1}^{n} (\frac{i}{n+1})^p$.

Algorithm	Throughput
Linked-list with fine-grained locks	$\frac{2p}{(n+1)\mathcal{L}_{cpu}}$
Flat-combining linked-list without combining	$\frac{2}{(n+1)\mathcal{L}_{cpu}}$
PIM-managed linked-list without combining	$\frac{2}{(n+1)\mathcal{L}_{pim}}$
Flat-combining linked-list with combining	$\frac{p}{(n-\mathcal{S}_p)\mathcal{L}_{cpu}}$
PIM-managed linked-list with combining	$\frac{p}{(n-\mathcal{S}_p)\mathcal{L}_{pim}}$

Table 1: Throughput of linked-lists

We calculate the throughput values in Table 1 in the following manner. In the linked-list with fine-grained locks, which has $(n+1)$ nodes including a dummy head node, each thread (CPU) executes its own operations to the linked-list. The key of a request is generated uniformly at random, so the average number of memory accesses by one thread for one operation is $(n+1)/2$ and hence the throughput of one thread is $2/((n+1)\mathcal{L}_{cpu})$. There are p threads running in parallel, so the total throughput is $2p/((n+1)\mathcal{L}_{cpu})$. The throughput of the flat-combining and the PIM-managed linked-lists without the combining optimization is calculated in a similar manner. For the flat-combining and the PIM-managed linked-lists with combining, it suffices to prove that the average number of memory accesses by a PIM core (or a combiner) batching and executing p random operation requests in one traversal is $n - \mathcal{S}_p$, which is essentially the expected number of pointers a PIM core (or a combiner) needs to go through to reach the position for the request with the largest key among the p requests. Note that we have ignored certain communication costs incurred in some linked-lists, such as the latency of a PIM core sending a result back to a waiting thread, and the latency of a combiner maintaining the combiner lock and the publication list in the flat-combining linked-list (we will discuss the publication list in more detail in Section 5), as they are negligible compared to the dominant costs of traversals over linked-lists.

It is easy to see that the PIM-managed linked-list with combining outperforms the linked-list with fine-grained locks, which is the best one among other linked-lists, if $\frac{\mathcal{L}_{cpu}}{\mathcal{L}_{pim}} = r_1 > \frac{2(n-\mathcal{S}_p)}{n+1}$. Given that $0 < \mathcal{S}_p \leq \frac{n}{2}$, the PIM-managed linked-list can outperform the linked-list with fine-grained locks as long as $r_1 \geq 2$. If we assume $r_1 = 3$, as estimated by prior work, the throughput of the PIM-managed linked-list with combining should be at least 1.5 times the throughput of the linked-list with fine-grained locks. Without combining, however, the PIM-managed linked-list *cannot* outperform the linked-list with fine-grained locks accessed by $p \geq r_1$ concurrent threads. On the other hand, the PIM-managed linked-list is expected to be r_1 times better than the flat-combining linked-list, with or without the combining optimization applied to both.

We implemented the linked-list with fine-grained locks and the flat-combining linked-list with and without the combining optimization. We tested them on a Dell server with 512 GB RAM and 56 cores on four Intel Xeon E7-4850v3 processors running at 2.2 GHz. To eliminate NUMA access effects, we ran experiments with only one processor, which is a NUMA node with 14 cores, a 35 MB shared L3 cache, and a private L2/L1 cache of size 256 KB/64 KB per core. Each core has 2 hyperthreads, for a total of 28 hyperthreads.

The throughput of each of the linked-lists, measured in operations per second, is presented in Figure 2. The results confirm the validity of our analysis in Table 1. The throughput of the flat-combining linked-list without the combining optimization is worse than the linked-list with fine-grained locks. Since the throughput of the flat-combining linked-list is a good indicator of the performance of the PIM-managed linked-list, we triple the throughput of the flat-combining linked-list to obtain the expected throughput of the PIM-managed linked-list, based on the assumption that $r_1 = 3$. As we can see, it is still below the throughput of the one with fined-grained locks. However, with the combining optimization, the performance of the flat-combining linked-list improves significantly and our PIM-managed linked-list with the combining optimization now outperforms all other data structures. We conclude that our PIM-managed linked-list is effective.

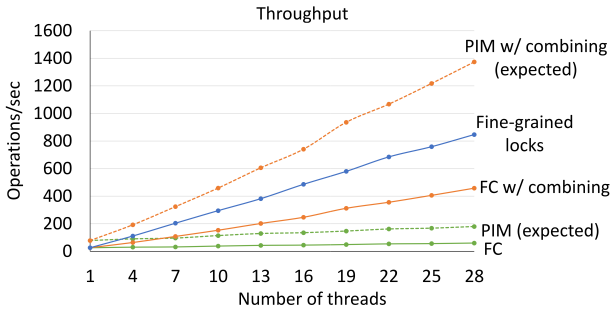

Figure 2: Experimental results of linked-lists. We evaluate the linked-list with fine-grained locks and the flat-combining linked-list (FC) with and without the combining optimization.

4.2 Skip-lists

Like the naive PIM-managed linked-list, the naive PIM-managed skip-list keeps the skip-list in a single vault and CPU cores send operation requests to the local PIM core that executes those operations. As we will see, this skip-list is less efficient than some existing skip-list algorithms.

Unfortunately, the combining optimization *cannot* be applied to skip-lists effectively. The reason is that for any two distant nodes in the skip-list, the paths threads must traverse to reach such nodes do *not* have large overlapping sub-paths. FloDB [7] uses a multi-insert operation for skip-lists, similar to the combining optimization we use for linked-lists. However, FloDB can ensure that the operations performed in a single traversal are close together because the operations are first grouped using a hash-table.

On the other hand, PIM memory usually consists of many vaults and PIM cores. For instance, the first generation of Hybrid Memory Cube [15] has up to 32 vaults. Hence, a PIM-managed skip-list can achieve much better performance if we can exploit the parallelism of multiple vaults. Here we present our PIM-managed skip-list with a *partitioning optimization*: A skip-list is divided into partitions of disjoint ranges of keys, stored in different vaults, so that a CPU sends its operation request to the PIM core of the vault to which the key of the operation belongs.

Figure 3 illustrates the structure of a PIM-managed skip-list. Each partition of a skip-list starts with a *sentinel node* which is a node with maximum height. For simplicity, assume that the max height H_{max} is predefined. A partition covers a key range between the key of its sentinel node and the key of the sentinel node of the next partition. CPUs also store a copy of each sentinel node in regular DRAM (see Figure 1) and this copy has an extra variable indicating the vault containing the sentinel node. The number of nodes with max height is very small with high probability, so the sentinel nodes can likely be found in the CPU caches because CPUs access them frequently.

When a CPU performs an operation for a key on the skip-list, it first compares the key with those of the sentinels, discovers which vault the key belongs to, and then sends its operation request to that vault's PIM core. After the PIM core retrieves the request, it executes the operation in the local vault and sends the result back to the CPU.

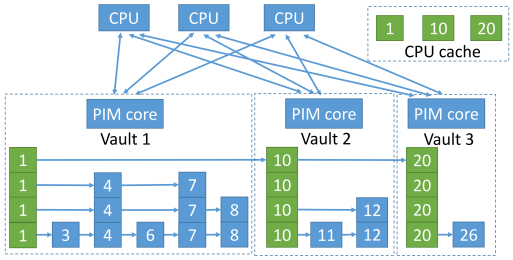

Figure 3: A PIM-managed skip-list with three partitions

We now discuss how we implement the PIM-managed skip-list when the key of each operation is an integer generated uniformly at random from range $[0, n]$ and the PIM memory has k vaults available. Initially we can create k partitions starting with fake sentinel nodes with keys 0, $1/k$, $2/k$,..., $(n-1)/k$, respectively, and allocate each partition in a different vault. The sentinel nodes are never deleted. If a new node to be added has the same key as a sentinel node, we insert it immediately after the sentinel node.

We compare the performance of our PIM-managed skip-list with k partitions to the performance of a flat-combining skip-list [25] and a lock-free skip-list [27], accessed concurrently by p CPUs. We also apply the partitioning optimization to the flat-combining skip-list, so that k combiners are in charge of k partitions of the skip-list. To simplify the comparison, we assume that all skip-lists have the same initial structure, i.e. skip-lists with partitions have extra sentinel nodes. We execute an equal number of add() and remove() requests, so that the size of the skip-list does not change dramatically. The keys of requests are generated uniformly at random.

The approximate throughput of each of these skip-lists is presented in Table 2, where β is the average number of nodes an operation has to access in order to find the location of its key in a skip-list ($\beta = \Theta(\log N)$, where N is the size of the skip-list). In the lock-free skip-list, p threads execute their own operations in parallel, so the throughput is roughly $p/(\beta \mathcal{L}_{cpu})$. Without the partitioning optimization, a combiner in the flat-combining skip-list and a PIM core in the PIM-managed skip-list both have to execute operations one by one sequentially, leading to throughput of roughly $\frac{1}{(\beta \mathcal{L}_{cpu})}$ and $\frac{1}{(\beta \mathcal{L}_{pim}+\mathcal{L}_{message})}$ respectively, where $\mathcal{L}_{message}$ is incurred by the PIM core sending a message with a result back to a CPU. After dividing these two skip-lists into k partitions, we can achieve a speedup of k for both of them, as k PIM cores and k combiners can serve requests in parallel now. Note that we have ignored certain costs in the lock-free skip-list and the two flat-combining skip-lists, such as the cost of a combiner's operations on the publication list in a flat-combining skip-list and the cost of CAS operations in the lock-free skip-list, so their actual performance could be even worse than what we show in Table 2.

Algorithm	**Throughput**
Lock-free skip-list	$\frac{p}{\beta \mathcal{L}_{cpu}}$
Flat-combining skip-list	$\frac{1}{\beta \mathcal{L}_{cpu}}$
PIM-managed skip-list	$\frac{1}{(\beta \mathcal{L}_{pim}+\mathcal{L}_{message})}$
Flat-combining skip-list with k partitions	$\frac{k}{\beta \mathcal{L}_{cpu}}$
PIM-managed skip-list with k partitions	$\frac{k}{(\beta \mathcal{L}_{pim}+\mathcal{L}_{message})}$

Table 2: Throughput of skip-lists

The results in Table 2 imply that the PIM-managed skip-list with k partitions is expected to outperform the second best skip-list, the lock-free skip-list, when $k > \frac{(\beta \mathcal{L}_{pim}+\mathcal{L}_{message})p}{\beta \mathcal{L}_{cpu}}$. Given that $\mathcal{L}_{message} = \mathcal{L}_{cpu} = r_1 \mathcal{L}_{pim}$ and $\beta = \Theta(\log N)$, $k > p/r_1$ should suffice. It is also easy to see that the performance of the PIM-managed skip-list is $\frac{\beta r_1}{\beta + r_1} \approx r_1$ times better than the flat-combining skip-list, when they have the same number of partitions.

Our experimental evaluation reveals similar results, as presented in Figure 4. We have implemented and run the flat-combining skip-list with different numbers of partitions and compared them with the lock-free skip-list. As the number of partitions increases, the performance of the flat-combining skip-list improves, attesting to the effectiveness of the partitioning optimization. Again, we believe the performance of the flat-combining skip-list is a good indicator of the performance of our PIM-managed skip-list. Therefore, according to the analytical results in Table 2, we can triple the throughput of a flat-combining skip-list to estimate the expected performance of a PIM-managed skip-list. As Figure 4 illustrates, when our PIM-managed skip-list has 8 or 16 partitions, it is expected to outperform the lock-free skip-list with up to 28 hardware threads.

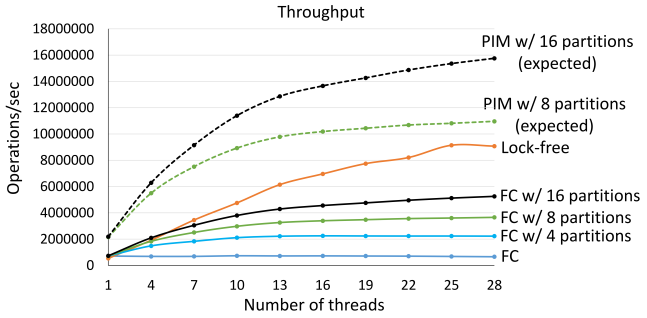

Figure 4: Experimental results of skip-lists. We evaluated the lock-free skip-list and the flat-combining skip-list (FC) with different numbers (1, 4, 8, 16) of partitions.

4.2.1 Skip-list Rebalancing. The PIM-managed skip-list performs well with a uniform distribution of requests. However, if the distribution of requests is *not* uniform, a static partitioning scheme will result in unbalanced partitions, with some PIM cores potentially being idle, while others having to serve a majority of the requests. To address this problem, we introduce a non-blocking protocol for migrating consecutive nodes from one vault to another.

The protocol works as follows. A PIM core p that manages a vault v' can send a message to another PIM core q, managing vault v, to request some nodes to be moved from v' to v. First, p sends a message notifying q of the start of the migration. Then p sends messages to q for adding those nodes into v one by one in ascending order according to the keys of the nodes. After all the nodes have been migrated, p sends notification messages to CPUs so that they can update their copies of sentinel nodes accordingly. After p receives acknowledgement messages from all CPUs, it notifies q of the end of migration. To keep the node migration protocol simple, we don't allow q to move those nodes to another vault again until p finishes its node migration.

During the node migration, p can still serve requests from CPUs. Assume that a request with key k_1 is sent to p when p is migrating nodes in a key range containing k_1. If p is about to migrate a node with key k_2 at the moment and $k_1 \geq k_2$, p serves the request itself. Otherwise, p must have migrated all nodes in the subset containing key k_1, and therefore p forwards the request to q which will serve the request and respond directly to the requesting CPU.

This skip-list is correct, because a request will eventually reach the vault that currently contains nodes in the key range that the request belongs to. If a request arrives to p which no longer holds the partition the request belongs to, p simply replies with a rejection to the CPU and the CPU will resend its request to the correct PIM core, because it has already updated its sentinels and knows which PIM core it should contact now.

Using this node migration protocol, the PIM-managed FIFO queue can support two rebalancing schemes: 1) If a partition has too many nodes, the local PIM core can move nodes in a key range to a vault that has fewer nodes; 2) If two consecutive partitions are both small, we can merge then by moving one to the vault containing the other.

In practice, we expect that rebalancing will not happen very frequently, so its overhead can be ameliorated by the improved efficiency resulting from the rebalanced partitions.

5 CONTENDED DATA STRUCTURES

In this section, we consider data structures that are often contended when accessed by many threads concurrently. In these data structures, operations compete for accessing one or more locations, creating a contention spot, which can become a performance bottleneck. Examples include head and tail pointers in queues and the top pointer of a stack.

These data structures have good locality; therefore, the contention spots are often found in shared CPU caches, such as the last-level cache in a multi-socket machine when shared by threads running on a single socket. Therefore, these data structures might seem to be a poor fit for near-memory computing: the advantage of faster memory access provided by PIM cannot be exercised because the frequently accessed data might stay in the CPU cache. However, such a perspective does not consider the overhead introduced by contention in a concurrent data structure where *many* threads access the *same* locations.

As a representative example of this class of data structures, we consider a FIFO queue, where concurrent enqueue and dequeue operations compete for the head and the tail of the queue, respectively. Although a naive PIM FIFO queue is not a good replacement for a well crafted concurrent FIFO queue, we show that, counterintuitively, PIM can still have benefits over a traditional concurrent FIFO queue. In particular, we exploit the *pipelining* of requests from CPUs, which can be done very efficiently in PIM, to design a PIM FIFO queue that can outperform state-of-the-art concurrent FIFO queues, such as the flat-combining FIFO queue [25] and the F&A FIFO queue [41].

5.1 FIFO queues

The structure of our PIM-managed FIFO queue is shown in Figure 5. A queue consists of a sequence of *segments*, each containing consecutive nodes of the queue. A segment is allocated in a PIM vault, with a head node and a tail node pointing to the first and the last nodes of the segment, respectively. A vault can contain multiple (likely non-consecutive) segments. There are two special segments—the *enqueue segment* and the *dequeue segment*. To enqueue a node, a CPU sends an enqueue request to the PIM core of the vault containing the enqueue segment. The PIM core then inserts the node to the head of the segment. Similarly, to dequeue a node, a CPU sends a dequeue request to the PIM core of the vault holding the dequeue segment. The PIM core then removes the node at the tail of the dequeue segment and sends the node back to the CPU.

Initially, the queue consists of an empty segment that acts as both the enqueue segment and the dequeue segment. When the length of the enqueue segment exceeds some threshold, the PIM core maintaining it notifies another PIM core to create a new segment as the new enqueue segment.[4] When the dequeue segment becomes empty and the queue has other segments, the dequeue segment

[4] Alternative designs where a CPU decides when to create new segments based on more complex criteria are also possible. We leave such designs as future work.

Algorithm 1 PIM-managed FIFO queue

```
1:  procedure enq(cid, u)
2:      if enqSeg == null then
3:          send message(cid, false);
4:      else
5:          if enqSeg.head ≠ null then
6:              enqSeg.head.next = u;
7:              enqSeg.head = u;
8:          else
9:              enqSeg.head = u;
10:             enqSeg.tail = u;
11:         enqSeg.count = enqSeg.count + 1;
12:         send message(cid, true);
13:         if enqSeg.count > threshold then
14:             cid′ = the CID of the PIM core chosen to maintain the new segment;
15:             send message(cid′, newEnqSeg());
16:             enqSeg.nextSegCid = cid′;
17:             enqSeg = null;

18: procedure newEnqSeg()
19:     enqSeg = new Segment();
20:     segQueue.enq(engSeg) ;
21:     notify the CPUs of the new enqueue segment;

22: procedure deq(cid)
23:     if deqSeg == null then
24:         send message(cid, false);
25:     else
26:         if deqSeg.tail ≠ null then
27:             send message(cid, deqSeg.tail);
28:             deqSeg.tail = deqSeg.tail.next;
29:         else
30:             if deqSeg == enqSeg then
31:                 send message(cid, null);
32:             else
33:                 send message(deqSeg.nextSegCid, newDeqSeg());
34:                 deqSeg = null;
35:                 send message(cid, false);

36: procedure newDeqSeg()
37:     deqSeg = segQueue.deq();
38:     notify the CPUs of the new dequeue segment;
```

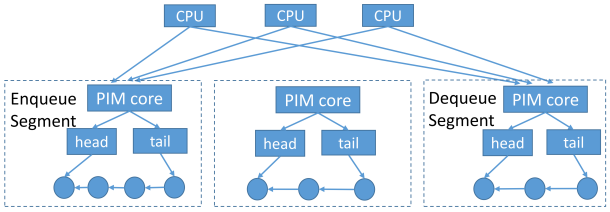

Figure 5: A PIM-managed FIFO queue with three segments

is deleted and the segment that was created first among all the remaining segments is designated as the new dequeue segment. This segment was created when the old dequeue segment acted as the enqueue segment and exceeded the length threshold. If the enqueue segment is different from the dequeue segment, enqueue and dequeue operations can be executed by two different PIM cores in parallel, improving the throughput. The F&A queue [41] also allows parallel enqueue and dequeue.

The pseudo-code of the PIM-managed FIFO queue is presented in Algorithm 1. Each PIM core has local variables *enqSeg* and *deqSeg* that are references to local enqueue and dequeue segments. When *enqSeg* (or *deqSeg*) is not null, it indicates that the PIM core is currently holding the enqueue (or dequeue) segment. Each PIM core also maintains a local queue *segQueue* for storing local segments. CPUs and PIM cores communicate via message(*cid*, *content*) calls, where *cid* is the unique core ID (CID) of the receiver and *content* is either a request or a response to a request.

Once a PIM core receives an *enqueue* request enq(*cid*, *u*) of node *u* from a CPU whose CID is *cid*, it first checks if it is holding the enqueue segment (line 2). If so, the PIM core enqueues *u* (lines 5-12), and otherwise sends back a message informing the CPU that the request is rejected (line 3) so that the CPU can resend its request to the right PIM core holding the enqueue segment (we will explain later how the CPU can find the right PIM core). After enqueuing *u*, the PIM core may find that the enqueue segment is longer than the threshold (line 13). If so, it sends a message with a newEnqSeg() request to the PIM core of another vault that is chosen to create a new enqueue segment. The PIM core then sets its *enqSeg* to null, indicating that it no longer deals with enqueue operations. Note that the CID *cid* of the PIM core chosen for creating the new segment is recorded in *enqSeg.nextSegCid* for future use in dequeue requests. As Procedure newEnqSeg() in Algorithm 1 shows, The PIM core receiving this newEnqSeg() request creates a new enqueue segment and enqueues the segment into its *segQueue* (lines 19-20). Finally, it notifies the CPUs of the new enqueue segment (we will discuss this notification in more detail later in this section).

Similarly, when a PIM core receives a *dequeue* request deq(*cid*) from a CPU with CID *cid*, it first checks whether it is holding the dequeue segment (line 23). If so, the PIM core dequeues a node and sends it back to the CPU (lines 26-28). Otherwise, it informs the CPU that this request has failed (line 24) and the CPU will have to resend its request to the right PIM core. If the dequeue segment is empty (line 29) and the dequeue segment is not the same as the enqueue segment (line 32), which implies that the FIFO queue is not empty, the PIM core sends a message with a newDeqSeg() request to the PIM core with CID *deqSeg.nextSegCid*. We know that this PIM core must hold the next segment, according to how we create new segments in enqueue operations, as shown in lines 14-16. Upon receiving the newDeqSeg() request, the PIM core retrieves from its *segQueue* the oldest segment it has created and makes it the new dequeue segment (line 37). Finally the PIM core notifies the CPUs that it is holding the new dequeue segment now.

We now explain how CPUs and PIM cores coordinate to make sure that the CPUs can find the right enqueue and dequeue segments, when their attempts fail due to enqueue/dequeue segment changes. We only discuss how to deal with enqueue segments, because the same methods can be applied to dequeue segments. A straightforward way to inform the CPUs is to have the owner PIM core of the new enqueue segment send notification messages to them (line 21) and wait until all the CPUs send back acknowledgment messages. However, if there is a slow CPU core that doesn't

reply in time, the PIM core has to wait for it and therefore other CPUs cannot have their requests executed. A more efficient, non-blocking method is to have the PIM core start serving new requests immediately after it has sent off the notifications to all CPUs. A CPU does not have to reply to those notifications in this case, but if its request later fails, it needs to send messages to all PIM cores to ask which PIM core is currently in charge of the enqueue segment. In either case, the correctness of the queue is guaranteed: at any time, there is only one enqueue segment and only one dequeue segment; only requests sent to them will be executed.

The PIM-managed FIFO queue can be further optimized. For example, the PIM core holding the enqueue segment can combine multiple pending enqueue requests and store the nodes to be enqueued in an array as a "fat" node of the queue, in order to reduce memory accesses. This optimization is also used in the flat-combining FIFO queue [25]. Even without this optimization, the PIM-managed FIFO queue still performs well, as we will show next.

5.2 Pipelining and Performance Analysis

We compare the performance of three concurrent FIFO queues—our PIM-managed FIFO queue, the flat-combining FIFO queue and the F&A-based FIFO queue [41]. The F&A-based FIFO queue is the most efficient concurrent FIFO queue we are aware of, where threads perform F&A operations on two shared variables, one for enqueues and the other for dequeues, to compete for slots in the FIFO queue to enqueue and dequeue nodes (see [41] for more details). The flat-combining FIFO queue we consider is based on the one proposed by [25], with a modification that threads compete for two "combiner locks", one for enqueues and the other for dequeues. We further simplify it based on the assumption that the queue is always non-empty, so that it doesn't have to deal with synchronization issues between enqueues and dequeues when the queue is empty. These assumptions give an advantage to the flat combining queue, to make it competitive with the two other queues, which can perform parallel enqueue and dequeue.

Let us first assume that a queue is long enough such that the PIM-managed FIFO queue has more than one segment, and enqueue and dequeue requests can be executed separately. Since enqueue/dequeue segment changes are infrequent, the overhead of such changes is negligible and therefore not included in our analysis. For example, if the threshold of segment length in line 13 of enq(*cid*, *u*) is a large integer n, then, in the worst case, changing an enqueue or dequeue segment happens only once every n requests. Moreover, a segment change only entails sending one message and a few steps of local computation. In our analysis, we focus on dequeue operations, because enqueues and dequeues are isolated from each other in all three FIFO queues when queues are long enough. The analysis of enqueues is similar.

Assume there are p concurrent dequeue requests by p threads. In the F&A queue, each thread needs to perform a F&A operation on a shared variable, serializing access to this shared variable. Therefore, the execution time of p requests is at least $p\mathcal{L}_{atomic}$. If we assume that each CPU makes a request immediately after its previous request completes, the throughput (per second) of the F&A queue is at most $\frac{1}{\mathcal{L}_{atomic}}$.

The flat-combining FIFO queue maintains a sequential FIFO queue and threads submit their requests into a *publication list*. The publication list consists of slots, one for each thread, to store their requests. After writing a request into the list, a thread competes with others for acquiring a lock to become the "*combiner*", which incurs one last-level cache access. The combiner then goes through the publication list to retrieve requests, executes operations for those requests, and writes results back to the list, while other threads with pending requests spin on their own slots, waiting for the results. The combiner therefore makes two last-level cache accesses[5] to each slot other than its own, one for reading the request and one for writing the result back. Thus, the execution time of p requests in this FIFO queue is at least $(2p-1)\mathcal{L}_{llc}$ and the throughput (per second) of this FIFO queue is at most $\frac{1}{2\mathcal{L}_{llc}}$ for large enough p.

Note that our analysis of the F&A-based and the flat-combining queues is performed in favor of them, as we consider only partial costs of their executions. We have ignored the latency of accessing and modifying queue nodes in the two FIFO queue algorithms. For dequeues, this latency can be high: nodes to be dequeued in a long queue are unlikely to be cached, so the combiner has to perform a sequence of memory accesses to dequeue them one by one. Moreover, the F&A-based queue may also suffer performance degradation under heavy contention, because contended F&A operations may perform worse in practice [16].

The performance of our PIM-managed FIFO queue seems poor at first sight: although a PIM core can update the queue efficiently, it takes a lot of time for the PIM core to send results back to CPUs one by one. To improve its performance, the PIM core can *pipeline* the execution of requests, as illustrated in Figure 6(a). Suppose p CPUs send p dequeue requests concurrently to the PIM core. The PIM core then retrieves a request from its message buffer (step 1 in the figure), dequeues a node (step 2) for the request, and sends the node back to the CPU (step 3). We can hide the message latency in step 3 as follows. After sending the message containing the node in step 3, the PIM core *immediately* retrieves the next request to execute, without blocking to wait for the previous message to arrive at its receiver. This way, the PIM core *pipelines* requests by overlapping the latency of message transfer in step 3 and the latency of memory accesses and local computations in steps 1 and 2 across multiple requests (see Figure 6(b)). Note that the PIM core still executes everything sequentially: it first sends the message for the current request before serving the next one.

The throughput of a PIM core is given by the costs of its memory accesses and local computations, as long as it has enough bandwidth to keep sending messages back to CPUs. In this FIFO queue algorithm, the PIM core sends a single small message per request, so bandwidth is unlikely to become a bottleneck.

Figure 6(b) illustrates that the execution time of p requests is the sum of the execution times of the first two steps for the p requests, plus the message transfer time of step 3 for the last request. During steps 1-2 of a dequeue, the PIM core only makes one memory access to read the node to be dequeued, and two L1 cache accesses to read and modify the tail node of the dequeue segment. Therefore, the total execution time of p requests, including the time $\mathcal{L}_{message}$ that

[5] We assume the combiner finds the slots in the last-level cache, to the benefit of the flat combining algorithm. If the slots are not found in cache, the cost will be higher, as the combiner will incur memory accesses instead.

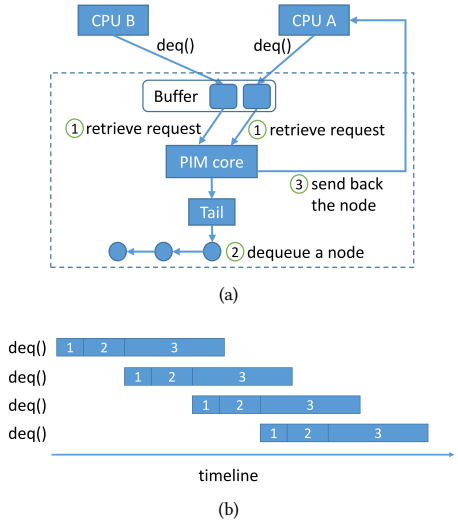

Figure 6: (a) The pipelining optimization, where a PIM core can start executing a new deq() (step 1 of deq() for CPU B), without waiting for the dequeued node of the previous deq() to return to CPU A (step 3). (b) The timeline of pipelining four deq() requests.

the CPUs spend in sending their requests to a PIM core concurrently at the beginning this execution, is $\mathcal{L}_{message} + p(\mathcal{L}_{pim} + \epsilon) + \mathcal{L}_{message}$, where $p(\mathcal{L}_{pim} + \epsilon)$ is the sum of the execution times of the first two steps for the p requests, and the second $\mathcal{L}_{message}$ is message transfer time of step 3 for the last request. ϵ is the total latency of the PIM core making two L1 cache accesses and sending one message. ϵ is negligible in our performance model.

Assume that each CPU makes another request immediately after it receives the result of its previous request and that there are enough (at least $2\mathcal{L}_{message}/\mathcal{L}_{pim}$) CPUs sending requests. We can prove that the PIM core can always find another request in its buffer after it executes one. Let x be the throughput of the PIM core in one second. By the same analysis as above, we have $\mathcal{L}_{message} + x(\mathcal{L}_{pim} + \epsilon) + \mathcal{L}_{message} = 1$, where 1 represents one second. Therefore, the throughput (per second) of the PIM-managed FIFO queue is approximately

$$x = \frac{1 - 2\mathcal{L}_{message}}{\mathcal{L}_{pim} + \epsilon} \approx \frac{1 - 2\mathcal{L}_{message}}{\mathcal{L}_{pim}} \approx \frac{1}{\mathcal{L}_{pim}},$$

since $\mathcal{L}_{message}$ is usually only hundreds of nanoseconds and much smaller than 1 (second).

Comparing the throughput values of the three FIFO queue algorithms, we can conclude that the PIM-managed FIFO queue with pipelining outperforms the other two FIFO queues when $2r_1/r_2 > 1$ and $r_1 r_3 > 1$. If we assume $r_1 = r_2 = 3$ and $r_3 = 1$, then the throughput of our PIM-managed FIFO queue is expected to be twice the throughput of the flat-combining queue and three times that of the F&A queue.[6]

[6] This does not imply that the F&A queue is faster than the flat combining queue, since we only consider part of the costs of these queues in our analysis.

When the PIM-managed FIFO queue is short, it may contain only one segment which deals with both enqueue and dequeue requests. In this case, its throughput is only half of the throughput shown above, but it is still at least as good as the throughput of the other two FIFO queues.

6 RELATED WORK

The PIM model is undergoing a renaissance. Studied for decades (e.g., [17, 20, 21, 32, 34, 44, 45, 51]), this model has recently re-emerged due to advances in 3D-stacked technology that can stack memory dies on top of a logic layer [8, 31, 33, 36, 40]. For example, a 3D-stacked memory prototype called the Hybrid Memory Cube [15] was recently released by industry, and the model has again become the focus of architectural research. Different PIM-based architectures have been proposed, either for general purpose workloads or for specific applications [1–6, 9, 13, 22, 23, 29, 30, 47–50, 53–55].

The PIM model has several advantages, including low energy consumption and high bandwidth (e.g., [1, 4, 53, 54]). Here, we focus on one more: low memory access latency [6, 23, 30, 40]. To our knowledge, we are the first to utilize PIM memory for designing efficient *concurrent data structures*. Although some researchers have studied how PIM memory can help speed up concurrent operations to data structures, such as parallel graph processing [1] and parallel pointer chasing on linked data structures [30], the applications they consider require very simple, if any, synchronization between operations. In contrast, operations to concurrent data structures can interleave in arbitrary orders, and therefore have to correctly synchronize with one another in all possible execution scenarios. This makes designing concurrent data structures with correctness guarantees, like linearizability [28], very challenging.

No prior work compares the performance of data structures in the PIM model with that of state-of-the-art concurrent data structures in the classical shared memory model. We analyze and evaluate concurrent linked-lists and skip-lists, as representatives of pointer-chasing data structures, and concurrent FIFO queues, as representatives of contended data structures. For linked-lists, we compare our PIM-managed implementation with well-known approaches such as fine-grained locking [24] and flat combining [18, 25, 26]. For skip-lists, we compare our implementation with a lock-free skip-list [27] and a skip-list with flat combining and the partitioning optimization. For FIFO queues, we compare our implementation with the flat-combining FIFO queue [25] and the F&A-based FIFO queue [41].

7 CONCLUSION

In this paper, we study how to design efficient data structures that can take advantage of the promising benefits offered by the Processing in Memory (PIM) paradigm. We analyze and compare the performance of our new PIM-managed data structures with traditional concurrent data structures that were proposed in the literature to take advantage of multiple processors. To this end, we develop a simplified performance model for PIM. Using this model, along with empirical performance measurements from a modern system, we show that naive PIM-managed data structures *cannot* outperform traditional concurrent data structures, due to the lack of parallelism and the high communication cost between the CPUs and

the PIM cores. To improve the performance of PIM data structures, we propose novel designs for low-contention pointer-chasing data structures, such as linked-lists and skip-lists, and for contended data structures, such as FIFO queues. We show that our new PIM-managed data structures can outperform state-of-the-art concurrent data structures, making PIM memory a promising platform for managing data structures. We conclude that it is very promising to examine novel data structure designs for the PIM paradigm, and hope future work builds upon our analyses to develop other types of PIM-managed data structures.

REFERENCES

[1] Junwhan Ahn, Sungpack Hong, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi. A scalable processing-in-memory accelerator for parallel graph processing. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, ISCA '15, pages 105–117, New York, NY, USA, 2015. ACM.

[2] Junwhan Ahn, Sungjoo Yoo, Onur Mutlu, and Kiyoung Choi. PIM-enabled instructions: A low-overhead, locality-aware processing-in-memory architecture. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, ISCA '15, pages 336–348, New York, NY, USA, 2015. ACM.

[3] Berkin Akin, Franz Franchetti, and James C. Hoe. Data reorganization in memory using 3D-stacked DRAM. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, ISCA '15, pages 131–143, New York, NY, USA, 2015. ACM.

[4] Erfan Azarkhish, Christoph Pfister, Davide Rossi, Igor Loi, and Luca Benini. Logic-base interconnect design for near memory computing in the Smart Memory Cube. *IEEE Trans. VLSI Syst.*, 25(1):210–223, 2017.

[5] Erfan Azarkhish, Davide Rossi, Igor Loi, and Luca Benini. High performance AXI-4.0 based interconnect for extensible Smart Memory Cubes. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, DATE '15, pages 1317–1322, San Jose, CA, USA, 2015. EDA Consortium.

[6] Erfan Azarkhish, Davide Rossi, Igor Loi, and Luca Benini. Design and evaluation of a processing-in-memory architecture for the Smart Memory Cube. In *Proceedings of the 29th International Conference on Architecture of Computing Systems – ARCS 2016 - Volume 9637*, pages 19–31, New York, NY, USA, 2016. Springer-Verlag New York, Inc.

[7] Oana Balmau, Rachid Guerraoui, Vasileios Trigonakis, and Igor Zablotchi. FloDB: Unlocking memory in persistent key-value stores. In *Proceedings of the Twelfth European Conference on Computer Systems*, EuroSys '17, pages 80–94, New York, NY, USA, 2017. ACM.

[8] Bryan Black, Murali Annavaram, Ned Brekelbaum, John DeVale, Lei Jiang, Gabriel H. Loh, Don McCaule, Pat Morrow, Donald W. Nelson, Daniel Pantuso, Paul Reed, Jeff Rupley, Sadasivan Shankar, John Shen, and Clair Webb. Die stacking (3D) microarchitecture. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 39, pages 469–479, Washington, DC, USA, 2006. IEEE Computer Society.

[9] Amirali Boroumand, Saugata Ghose, Brandon Lucia, Kevin Hsieh, Krishna Malladi, Hongzhong Zheng, and Onur Mutlu. LazyPIM: An efficient cache coherence mechanism for processing-in-memory. *IEEE Computer Architecture Letters*, 2016.

[10] Irina Calciu, Siddhartha Sen, Mahesh Balakrishnan, and Marcos K. Aguilera. Black-box concurrent data structures for NUMA architectures. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '17, pages 207–221, New York, NY, USA, 2017. ACM.

[11] Kevin K. Chang. *Understanding and Improving Latency of DRAM-Based Memory Systems.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2017.

[12] Kevin K. Chang, Abhijith Kashyap, Hasan Hassan, Saugata Ghose, Kevin Hsieh, Donghyuk Lee, Tianshi Li, Gennady Pekhimenko, Samira Khan, and Onur Mutlu. Understanding latency variation in modern DRAM chips: Experimental characterization, analysis, and optimization. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, SIGMETRICS '16, pages 323–336, New York, NY, USA, 2016. ACM.

[13] Kevin K. Chang, Prashant J. Nair, Donghyuk Lee, Saugata Ghose, Moinuddin K. Qureshi, and Onur Mutlu. Low-cost inter-linked subarrays (LISA): enabling fast inter-subarray data movement in DRAM. In *IEEE International Symposium on High Performance Computer Architecture, HPCA 2016, Barcelona, Spain, March 12-16, 2016*, pages 568–580, 2016.

[14] Kevin K. Chang, A. Giray Yaglikci, Saugata Ghose, Aditya Agrawal, Niladrish Chatterjee, Abhijith Kashyap, Donghyuk Lee, Mike O'Connor, Hasan Hassan, and Onur Mutlu. Understanding reduced-voltage operation in modern dram devices: Experimental characterization, analysis, and mechanisms. In *to appear in Proceedings of the 2017 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, SIGMETRICS '17.

[15] Hybrid Memory Cube Consortium. Hybrid Memory Cube specification 1.0, 2013.

[16] Tudor David, Rachid Guerraoui, and Vasileios Trigonakis. Everything you always wanted to know about synchronization but were afraid to ask. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, SOSP '13, pages 33–48, New York, NY, USA, 2013. ACM.

[17] Duncan G. Elliott, W. Martin Snelgrove, and Michael Stumm. Computational RAM: A memory-SIMD hybrid and its application to DSP. In *Proceedings of the IEEE 1992 Custom Integrated Circuits Conference*, CICC '92, pages 30.6.1–30.6.4, Piscataway, NJ, USA, 1992. IEEE Press.

[18] Panagiota Fatourou and Nikolaos D. Kallimanis. Revisiting the combining synchronization technique. In *Proceedings of the 17th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP '12, pages 257–266, New York, NY, USA, 2012. ACM.

[19] Keir Fraser. Practical lock-freedom. Technical Report UCAM-CL-TR-579, University of Cambridge, Computer Laboratory, February 2004.

[20] Maya Gokhale, Bill Holmes, and Ken Iobst. Processing in memory: The Terasys massively parallel PIM array. *Computer*, 28(4):23–31, April 1995.

[21] Mary Hall, Peter Kogge, Jeff Koller, Pedro Diniz, Jacqueline Chame, Jeff Draper, Jeff LaCoss, John Granacki, Jay Brockman, Apoorv Srivastava, William Athas, Vincent Freeh, Jaewook Shin, and Joonseok Park. Mapping irregular applications to DIVA, a PIM-based data-intensive architecture. In *Proceedings of the 1999 ACM/IEEE Conference on Supercomputing*, SC '99, New York, NY, USA, 1999. ACM.

[22] M. Hashemi, O. Mutlu, and Y. N. Patt. Continuous Runahead: Transparent hardware acceleration for memory intensive workloads. In *Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '16, Oct 2016.

[23] Milad Hashemi, Khubaib, Eiman Ebrahimi, Onur Mutlu, and Yale N. Patt. Accelerating dependent cache misses with an enhanced memory controller. In *Proceedings of the 43rd International Symposium on Computer Architecture*, ISCA '16, pages 444–455, Piscataway, NJ, USA, 2016. IEEE Press.

[24] Steve Heller, Maurice Herlihy, Victor Luchangco, Mark Moir, William N. Scherer, and Nir Shavit. A lazy concurrent list-based set algorithm. In *Proceedings of the 9th International Conference on Principles of Distributed Systems*, OPODIS'05, pages 3–16, Berlin, Heidelberg, 2006. Springer-Verlag.

[25] Danny Hendler, Itai Incze, Nir Shavit, and Moran Tzafrir. Flat combining and the synchronization-parallelism tradeoff. In *Proceedings of the Twenty-second Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '10, pages 355–364, New York, NY, USA, 2010. ACM.

[26] Danny Hendler, Itai Incze, Nir Shavit, and Moran Tzafrir. Scalable flat-combining based synchronous queues. In *Proceedings of the 24th International Conference on Distributed Computing*, DISC'10, pages 79–93, Berlin, Heidelberg, 2010. Springer-Verlag.

[27] Maurice Herlihy and Nir Shavit. *The Art of Multiprocessor Programming*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[28] Maurice P. Herlihy and Jeannette M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, July 1990.

[29] Kevin Hsieh, Eiman Ebrahimi, Gwangsun Kim, Niladrish Chatterjee, Mike O'Connor, Nandita Vijaykumar, Onur Mutlu, and Stephen W. Keckler. Transparent offloading and mapping (TOM): Enabling programmer-transparent near-data processing in GPU systems. In *Proceedings of the 43rd International Symposium on Computer Architecture*, ISCA '16, pages 204–216, Piscataway, NJ, USA, 2016. IEEE Press.

[30] Kevin Hsieh, Samira Khan, Nandita Vijaykumar, Kevin K Chang, Amirali Boroumand, Saugata Ghose, and Onur Mutlu. Accelerating pointer chasing in 3D-stacked memory: Challenges, mechanisms, evaluation. In *IEEE 34th International Conference on Computer Design, ICCD 2016*, pages 25–32. IEEE, 2016.

[31] Joe Jeddeloh and Brent Keeth. Hybrid memory cube new DRAM architecture increases density and performance. In *Symposium on VLSI Technology, VLSIT 2012*, pages 87–88. IEEE, 2012.

[32] Yi Kang, Wei Huang, Seung-Moon Yoo, Diana Keen, Zhenzhou Ge, Vinh Vi Lam, Josep Torrellas, and Pratap Pattnaik. FlexRAM: Toward an advanced intelligent memory system. In *Proceedings of the IEEE International Conference On Computer Design*, ICCD '99.

[33] Joonyoung Kim and Younsu Kim. HBM: Memory solution for bandwidth-hungry processors. *2014 IEEE Hot Chips 26 Symposium (HCS)*, 00:1–24, 2014.

[34] Peter M. Kogge. EXECUBE-a new architecture for scaleable MPPs. In *Proceedings of the 1994 International Conference on Parallel Processing - Volume 01*, ICPP '94, pages 77–84, Washington, DC, USA, 1994. IEEE Computer Society.

[35] Donghyuk Lee. *Reducing DRAM Latency at Low Cost by Exploiting Heterogeneity.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2017.

[36] Donghyuk Lee, Saugata Ghose, Gennady Pekhimenko, Samira Khan, and Onur Mutlu. Simultaneous multi-layer access: Improving 3D-stacked memory bandwidth at low cost. *ACM Trans. Archit. Code Optim.*, 12(4):63:1–63:29, January 2016.

[37] Donghyuk Lee, Samira Khan, Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungnirun, Gennady Pekhimenko, Vivek Seshadri, and Onur Mutlu.

Design-induced latency variation in modern dram chips: Characterization, analysis, and latency reduction mechanisms. In *to appear in Proceedings of the 2017 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, SIGMETRICS '17.

[38] Donghyuk Lee, Yoongu Kim, Gennady Pekhimenko, Samira Manabi Khan, Vivek Seshadri, Kevin Kai-Wei Chang, and Onur Mutlu. Adaptive-latency DRAM: optimizing DRAM timing for the common-case. In *21st IEEE International Symposium on High Performance Computer Architecture, HPCA 2015, Burlingame, CA, USA, February 7-11, 2015*, pages 489–501, 2015.

[39] Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, and Onur Mutlu. Tiered-latency DRAM: A low latency and low cost DRAM architecture. In *19th IEEE International Symposium on High Performance Computer Architecture, HPCA 2013, Shenzhen, China, February 23-27, 2013*, pages 615–626, 2013.

[40] Gabriel H. Loh. 3D-stacked memory architectures for multi-core processors. In *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ISCA '08, pages 453–464, Washington, DC, USA, 2008. IEEE Computer Society.

[41] Adam Morrison and Yehuda Afek. Fast concurrent queues for x86 processors. In *Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP '13, pages 103–112, New York, NY, USA, 2013. ACM.

[42] Onur Mutlu. Memory scaling: a systems architecture perspective. In *Proceedings of the 5th International Memory Workshop*, IMW '13, 2013.

[43] Onur Mutlu and Lavanya Subramanian. Research problems and opportunities in memory systems. *Supercomputing Frontiers and Innovations*, 1, 2014.

[44] Mark Oskin, Frederic T. Chong, and Timothy Sherwood. Active pages: A computation model for intelligent memory. In *Proceedings of the 25th Annual International Symposium on Computer Architecture*, ISCA '98, pages 192–203, Washington, DC, USA, 1998. IEEE Computer Society.

[45] David Patterson, Thomas Anderson, Neal Cardwell, Richard Fromm, Kimberly Keeton, Christoforos Kozyrakis, Randi Thomas, and Katherine Yelick. A case for intelligent RAM. *IEEE Micro*, 17(2):34–44, March 1997.

[46] W. Pugh. Concurrent maintenance of skip lists. Technical report, University of Maryland at College Park, 1990.

[47] Vivek Seshadri, Kevin Hsieh, Amirali Boroumand, Donghyuk Lee, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry. Fast bulk bitwise AND and OR in DRAM. *IEEE Comput. Archit. Lett.*, 14(2):127–131, July 2015.

[48] Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. RowClone: Fast and energy-efficient in-DRAM bulk data copy and initialization. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-46, pages 185–197, New York, NY, USA, 2013. ACM.

[49] Vivek Seshadri, Donghyuk Lee, Thomas Mullins, Hasan Hassan, Amirali Boroumand, Jeremie Kim, Michael A. Kozuch, Onur Mutlu, Phillip B. Gibbons, and Todd C. Mowry. Buddy-ram: Improving the performance and efficiency of bulk bitwise operations using DRAM. *CoRR*, abs/1611.09988, 2016.

[50] Vivek Seshadri and Onur Mutlu. The processing using memory paradigm: In-DRAM bulk copy, initialization, bitwise AND and OR. *CoRR*, abs/1610.09603, 2016.

[51] Harold S. Stone. A logic-in-memory computer. *IEEE Trans. Comput.*, 19(1):73–78, January 1970.

[52] J. Valois. *Lock-free Data Structures*. PhD thesis, Rensselaer Polytechnic Institute, Troy, NY, USA, 1996.

[53] Dongping Zhang, Nuwan Jayasena, Alexander Lyashevsky, Joseph L. Greathouse, Lifan Xu, and Michael Ignatowski. TOP-PIM: Throughput-oriented programmable processing in memory. In *Proceedings of the 23rd International Symposium on High-performance Parallel and Distributed Computing*, HPDC '14, pages 85–98, New York, NY, USA, 2014. ACM.

[54] Qiuling Zhu, Berkin Akin, H. Ekin Sumbul, Fazle Sadi, James C. Hoe, Larry T. Pileggi, and Franz Franchetti. A 3D-stacked logic-in-memory accelerator for application-specific data intensive computing. In *IEEE International 3D Systems Integration Conference, 3DIC 2013, San Francisco, CA, USA, October 2-4, 2013*, pages 1–7, 2013.

[55] Qiuling Zhu, Tobias Graf, H. Ekin Sumbul, Larry T. Pileggi, and Franz Franchetti. Accelerating sparse matrix-matrix multiplication with 3D-stacked logic-in-memory hardware. In *IEEE High Performance Extreme Computing Conference, HPEC 2013, Waltham, MA, USA, September 10-12, 2013*, pages 1–6, 2013.