Dynamically Adapting Page Migration Policies Based on Applications' Memory Access Behaviors

SHASHANK ADAVALLY, MAHZABEEN ISLAM, and KRISHNA KAVI,
University of North Texas

There have been numerous studies on heterogeneous memory systems comprised of faster DRAM (e.g., 3D stacked HBM or HMC) and slower non-volatile memories (e.g., PCM, STT-RAM). However, most of these studies focused on static policies for managing data placement and migration among the different memory devices. These policies are based on the average behavior across a range of applications. Results show that these techniques do not always result in higher performance when compared to systems that do not migrate data across the devices: some applications show performance gains, but other applications show performance losses. It is possible to utilize offline analyses to identify which applications benefit from page migration (migration friendly) and use page migration only with those applications. However, we observed that several applications exhibit both migration friendly and migration unfriendly behaviors during different phases of execution supporting a need for adaptive page migration techniques. We introduce and evaluate techniques that dynamically adapt to the behavior of applications and either reduce or increase migrations, or even halt migrations. Our adaptive techniques show performance gains for both migration friendly (on average of 81% over no migrations) and unfriendly workloads (by an average of 3%): it should be remembered that previous migration techniques resulted in performance losses for unfriendly workloads.

CCS Concepts: • **Computer systems organization → Heterogeneous (hybrid) systems**;

Additional Key Words and Phrases: Heterogeneous memory systems, flat address memory, dynamic page migration, reverse migration

ACM Reference format:
Shashank Adavally, Mahzabeen Islam, and Krishna Kavi. 2021. Dynamically Adapting Page Migration Policies Based on Applications' Memory Access Behaviors. *J. Emerg. Technol. Comput. Syst.* 17, 2, Article 16 (March 2021), 24 pages.
https://doi.org/10.1145/3444750

1 INTRODUCTION

High performance applications and emerging data-centric applications need memory systems with very large capacities (hundreds of gigabytes to terabytes), high bandwidth, and energy efficiency [3, 16]. For example, SAP HANA in-memory database system requires 256 GB to multiple terabytes per host [3], and Spark in-memory analytics provides higher performance when run

This research was supported in part by NSF awards 1828105 and 1361806.
Authors' address: S. Adavally, M. Islam, and K. Kavi, University of North Texas, 1155 Union Circle, Denton, TX, 76203; emails: {ShashankAdavally, MahzabeenIslam}@my.unt.com, Krishna.Kavi@unt.edu.

1550-4832/2021/03-ART16 $15.00
https://doi.org/10.1145/3444750

with 12 TB of memory [16]. In addition, the value of these large amounts of gathered data depends on how fast the data can be analyzed to make decisions [16].

The architecture community has reacted to these needs with new memory technologies including 3D stacked DRAMs and very dense non-volatile memories (NVMs). Different organizations for combining these diverse memory technologies into a system architecture have been investigated, including hierarchical organizations (i.e., using 3D DRAM as last level cache (LLC)) or flat-address memories where 3D DRAM (e.g., HBM [9], HMC [7]), DDR, and NVM devices (e.g., STT-RAM [20], phase change memories (PCMs) [28]) form a single memory address space. In such flat-address systems, to effectively reduce average memory access times, heavily accessed pages (hot pages) are migrated from slower memories (NVM) to faster memories (3D DRAM); cold pages are moved from faster memories to slower memories to make room for the hot pages. Pages can be migrated (or swapped) at regular intervals (epoch based) or individually (on-the-fly). The migration of pages between the memory systems incur execution and energy overheads. In addition to the cost of actual data movement between memory devices, OS tables (translation look-aside buffers (TLBs), page tables) must also be updated since physical addresses (PAs) in such memory systems are based on the physical location of pages and a migration changes PAs: we call this process of changing PAs and updating system tables address reconciliation (AR).

Numerous designs have evaluated the efficiencies of different page migration techniques. One thing is clear from these studies: no single approach leads to consistent performance improvement for all applications. Some applications may actually see a performance degradation due to page migrations [21, 27]. It may be possible to perform offline analysis of memory access behaviors of applications and categorize them as migration friendly (applications that show performance gains from page migration—for these applications, performance gains outweigh migration overheads) and migration unfriendly (applications that do not show performance gains—overheads outweigh performance gains from migration), and use page migrations for only the migration friendly applications. However, offline analyses are often coarse grained and may not capture evolving behaviors of applications: applications may have both migration friendly and unfriendly phases. It is necessary to design adaptive migration techniques to dynamically increases or reduces migrations and even turns off migrations to adapt to changing behaviors of applications. We propose such adaptive page migrations techniques in this article. The key contributions of our work are as follows:

—*Adaptive migration polices*: Previous page migration techniques relied on fixed hotness thresholds: a page is migrated from slow memories to faster memories when the number of times that page was accessed exceeds the hotness threshold. In contrast, we control page migration policies based on applications' memory access behaviors. Our technique increases or reduces the hotness thresholds to reduce or increase the number of pages migrated based on either the number of pages migrated over a window of observation or based on the observed benefits of page migrations (were pages accessed after the migration to faster memories).

—*AR overheads can defeat the benefits of page migration*: To eliminate AR, we explore the benefit of reverse migrating pages to their original locations, particularly when the migrated pages are no longer heavily accessed. Reverse migration makes page migration invisible to the OS. However, reverse migrations can result in excessive data movement between slow and fast memories. In this work, we evaluate the effectiveness of the reverse migration technique.

Epoch-based approaches migrate "hot" pages at the end of an epoch. In such systems, some hot pages may have exhausted their usefulness by the time they are migrated. Our previous research

[14] and the MemPod study [27] have observed that migrating recently accessed hot pages results in better performance than migrating the "hottest" pages with accesses accrued over a longer period (i.e., an epoch). We migrate pages as soon as they become hot (we call this on-the-fly (OTF) migration). Epoch-based approaches rely on the OS for AR, which can be excessive. In our previous study [14], we used a special hardware migration controller (MigC) that includes a small remap table (or ReMap table in our system) to track physical locations of recently migrated pages to aid in redirecting accesses to correct page locations. Periodically, older entries in the ReMap table are deleted, after MigC updates page table entries (PTEs) and TLBs, making room for new page migrations (i.e., after AR). We based our adaptive migration techniques and reverse migration techniques on our previous design. We extended MigC to monitor applications' memory access behaviors to dynamically adapt page migrations.

The rest of the article is organized as follows. Section 2 includes the motivation for adaptive page migration techniques. We also include a description of our MigC, as well as the migration and AR processes. We include this information (previously reported in our previous work [14]) to make this contribution self-contained. Section 3 describes our adaptive migration techniques and the reverse migration technique. Section 4 contains our experimental setup and the benchmarks used for evaluation. Section 5 includes an analysis of the results from our experiments. We include both performance and energy results when using our adaptive migration policies. Section 6 includes a discussion of research that is closely related to ours, and Section 7 summarizes the conclusions of this study and further research that can be explored.

2 BACKGROUND AND MOTIVATION

A number of heterogeneous memory investigations (e.g., [21, 27, 35]) and our previous study [14] show that not all applications benefit from page migrations since page migrations incur performance overheads due to extra data movement, as well as overheads for AR. It is possible to develop offline analyses to categorize applications as migration friendly (applications that show performance gains) and migration unfriendly (applications that show performance losses) so that page migration is enabled only for migration friendly workloads. In our previous work [14], we developed one such offline classification by analyzing memory accesses to main memory pages (when the accesses miss the cache hierarchy). We then created a histogram that shows how many pages received a certain number of accesses. We discovered that an exponential-shaped histogram indicates that very few pages receive most accesses and that those applications benefit by either placing those few pages in the faster (HBM) memory at the start of execution, or migrated to HBM on demand. This is the case with mcf (one of the benchmarks) where just 3% of all pages cause 97% of memory accesses.[1] Thus, mcf gains significant performance from most page migration policies, and it is classified as a migration friendly application. However, applications exhibiting uniform-shaped histograms indicate that most or all pages receive about the same number of accesses, implying that too many pages may be migrated if a fixed hotness threshold is used for migrating pages, and the migration overheads outweigh performance gains. This behavior is exhibited by milc (another one of our benchmarks): 65% of pages contribute to 82% of all accesses, and this application does not benefit from page migration and will be classified as a migration unfriendly workload.

We also tracked the usefulness of pages that were recently migrated to faster memory. Migration of pages to faster memories results in performance gains if those pages continue to be heavily used, because these accesses will be satisfied by faster memories. We defined the migration benefit

[1]We omit details of the analysis and our approach for classifying applications as migration friendly or not friendly since offline analysis is not a key contribution of this work and that information was published in our earlier work [14].

Table 1. Migration Friendliness of Applications

Friendliness	Benchmark
Very friendly	mcf, mix1, mix2, mix4, mix5
Moderately friendly	lbm, omnetpp, astar cactus, BFS, mix3, mix6
Least friendly or unfriendly	milc, gems, zeusmp xalanc, braves, miniFE lulesh, xsbench, CoMD

quotient (MBQ) to measure the average usefulness of recently migrated pages. Relying on the histograms and MBQ, we classified applications as very (migration) friendly, moderately friendly, and migration unfriendly. Table 1 shows a possible classification of our benchmarks based on such an analysis.

2.1 On-the-Fly Page Migration

In this section, we will include a brief description of our previous work [14] to provide sufficient details needed to understand the contribution of this work. This section also provides the motivation for our adaptive migration techniques.

Unlike approaches that rely on epochs (e.g., 10-ms intervals) to track page access counts to determine which hot pages to migrate, we migrate a page as soon it receives a certain number of accesses (hotness threshold). We call this the *OTF migration technique*. OTF migration performs better than epoch-based page migration techniques since we migrate recent hot pages [14]. This is in line with the observations made by Prodromou et al. [27] that migrating recently accessed hot pages results in better performance than migrating the "hottest" pages with accesses accrued over a longer period. Since in OTF migration a page migration can take place at any time, it is important to ensure that a migration does not halt user program execution.[2] Moreover, OS-based AR on each page migration is prohibitive for OTF migration. To mitigate these issues, we devised a special hardware called *MigC*, placed on the processor chip, which performs actions necessary for our OTF page migration.

Figure 1 shows a high-level system architecture of MigC. There are hot and cold buffers to temporarily store data from pages as they are being migrated. There is a Wait queue in MigC, which holds read/write requests from LLC for the currently migrating pages; these requests will be serviced from the hot/cold buffers. The memory controllers (MCs) are equipped with a separate Migration queues (Mig.Q) to service requests from MigC for the migrating pages. There is a small ReMap table that holds new physical page addresses of the migrated pages. The ReMap table is consulted on every LLC miss (or on a write-back) using the old PA to find the new location. The size of the ReMap table is kept small (e.g., 1024 entries) so that it can be placed on-chip. Whenever the table is full to a certain level, say 50%, the AR process starts—that is, entries from the ReMap table are deleted and the new PAs are made visible to the OS as discussed in Section 2.3 and in our previous work [14].

[2]Epoch-based approaches stop program execution and migrate several hot pages at the end of an epoch. The OS manages the migration and updates TLBs and PTEs with new PAs.

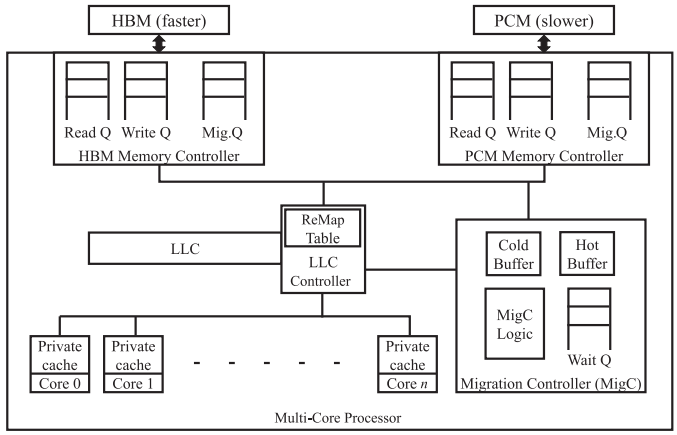

Fig. 1. High-level system architecture.

2.2 User Transparent Page Migration

We migrate a hot page from slower PCM into a free frame of HBM if available (one-way migration), or select a HBM page that has not been accessed recently (LRU cold page) and swap the hot and cold pages (two-way migration). Consider that MigC finds that a PCM page (say page A with PA, PA 8192) meets the hotness threshold for migration. MigC then finds a cold page from HBM (say page B with PA, PA 0) to swap with the hot page. MigC inserts entries for pages A and B in the ReMap table with their current OS visible PAs (namely 8192 and 0) and future PA (after migration, namely 0 and 8192). The ReMap table is always looked up using an OS visible (original) PA. Mig flag is set to 1 when these pages are being migrated and a Pair flag is set to 1 to indicate a two-way migration involving a hot and cold pair (this flag will be set to 0 for one-way migration). The Pair flag will be checked during AR to update PTEs for both (or one) pages involved in the migration.

MigC waits for any pending read requests to the pages involved in the migration that were already issued to complete. Then, MigC starts reading hot and cold pages into their respective buffers (inside MigC). Any new requests (after the migration is initiated) for these pages from LLC will be held in the MigC resident Wait queue and will be served from these buffers. After completely reading the migrating page contents into the buffers, MigC starts writing contents of buffers to their respective new page frames. When migrations are completed, the Mig flag for these pages will be reset (to indicate completion of migration). All future requests for these pages will be directed to proper new locations based on the ReMap table information.

2.3 Address Reconciliation

AR can be eliminated (and make page migration transparent to the OS) by using very large ReMap tables, sufficient to track all migrated pages during the lifetime of an application. However, this is not practical for emerging systems with very large memories (several hundred gigabytes to terabytes). We use a very small ReMap table and periodically evict old entries to make room for new entries for future migrations. Removing entries from the ReMap table requires updates to PAs (i.e., AR) to reflect the new location of the page consistently throughout the system, and making the new PAs visible to the OS. We reconcile entries from the ReMap table pairwise if the Pair flag is 1, thus updating the PAs of the pages swapped during the migration. When the Pair flag is 0, then we perform AR only for that entry. The following actions must be performed to ensure correct AR. We use the same example hot and cold page pair, A (PA = 8192) and B (PA = 0), respectively. First,

all cache lines from these pages, which are currently residing in the cache hierarchies and tagged with the OS visible (old) PA, must be invalidated (and dirty lines written back), since the current OS visible PA will be replaced with the new PA. All future accesses to these pages will only have access to the new PA. Next, corresponding PTEs for A and B need to be updated with new PAs. The TLB entries in all cores using the old PA must also be invalidated (known as TLB shootdown).

2.3.1 AR: OS vs. Hardware. Linux[3] performs the following functions when the virtual to PA mapping of a page is changed:

(i) *flush_cache_page()*,
(ii) change PTE,
(iii) *flush_tlb_page()* [23].

The function *flush_cache_page()* takes necessary parameters (a pointer to the process address space, the virtual address (VA) and the associated page frame number) and writes back any dirty cache lines of that page to memory and invalidates the cache lines belonging to that page. This process halts the user program resulting in large overhead. We found that on average it takes 4 μs to flush cache lines of a page using CLFLUSH x86 instruction on a processor running at 2.26 GHz. To update PTE, Linux acquires page table lock and changes PTE and also executes *flush_tlb_page()* to invalidate all TLBs (TLB shootdown) with old VA to PA translation. The OS releases the lock upon completing these actions. The TLB shootdown is costly because it uses IPI (interprocess interrupt) to invalidate TLB entries in every core that contains an entry with old PA. The delay grows non-linearly with number of cores [2, 30, 36]. As reported by Meswani et al. [21], TLB shootdown may take up to 4, 5, 8, and 13 μs for 4, 8, 16, and 32 cores, respectively, on an AMD 32-core system running Linux.

In our hardware-based approach, we configure MigC as a pseudo-processor that can send "write invalidate" requests over the coherency network for each of the cache lines of the pages under reconciliation, requiring all caches to write-back any dirty lines to memory and invalidate their cache lines for these pages (instead of CLFLUSH instruction). MigC will be configured such that it can send coherence requests to other caches and receive acknowledgments back from them; however, other caches will never send requests to or wait for any acknowledgments from MigC. For TLB shootdown, we rely on a shared TLB directory that contains all private TLB entries along with process identifiers (i.e., address space identifier) and core residency information. MigC initiates TLB shootdown by sending the associated VAs to the shared TLB directory. The shared TLB directory then maps these VAs to necessary entries and requests cores to invalidate these TLB entries somewhat similar to that used in the work of Villavieja et al. [36]. We envision that actual invalidation at each core will be carried out by a per-core hardware invalidation controller without interrupting the core, and the upper bound of time required for completing such invalidations is assumed to be a round-trip off-chip memory access latency [36].

2.3.2 One Final Issue in AR. To update PTE to reflect the new PA of a migrated page and invalidate associated TLB entries we need the VA of the page. However, our ReMap table contains only the original PA of a page and not its VA. Moreover, since the same page can be shared by multiple processes and each process may have a different VA corresponding to the PA of the page, we need to obtain all possible VAs. Linux keeps descriptors for every allocated physical page frame that maintains bookkeeping information on the number of PTEs referring to this page frame and pointers to such PTEs [5]. By using existing Linux reverse mapping function, we can obtain the

[3]We use Linux-based systems in all of our experiments, which makes it easier to compare our results with those reported in the literature.

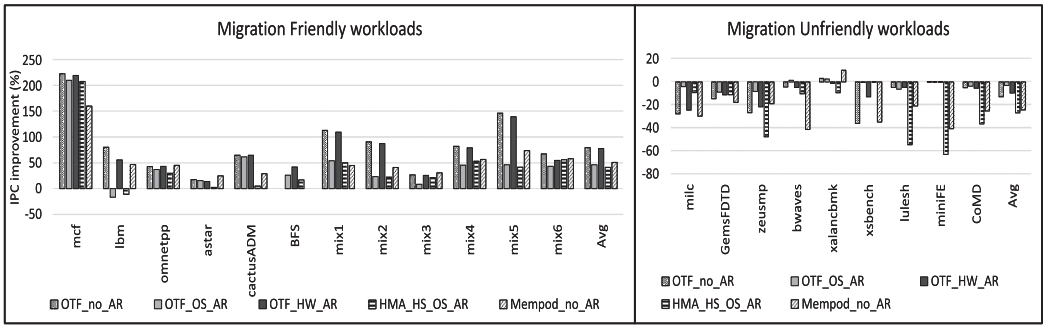

Fig. 2. IPC improvement (%) of different page migration and AR policies over a no-migration baseline with static hotness thresholds (negative *y*-axis shows degradation).

list of PAs of PTEs that hold mappings to this specific page frame number and associated VAs with ASIDs [5]. We account for all delays involved for these OS functions needed to implement our AR and page migrations using previously reported numbers and actual experimental data on a real system (see Table 3). More details on how our MigC performs AR can be found in our previous work [14].

2.4 Analysis of Fixed Hotness Threshold Experiments

As a motivation for the adaptive migration techniques presented in this article, we reproduced some results from our previous work [14]. We include results using OTF migrations without any AR (labeled as OTF_no_AR), assuming a sufficiently large on-chip ReMap table to track all pages migrated during program executions. This is unrealistic but provides a data point for comparison. We also included OTF with OS-based AR (labeled as OTF_OS_AR) and OTF with our hardware-based AR (labeled as OTF_HW_AR). This allows us to directly compare the benefits of using hardware instead of the OS for AR. We compared our OTF schemes with an epoch-based page migration study [21], which uses a combination of hardware and OS for AR (we refer to this as HMA_HS_OS_AR), and with MemPod [27] with no AR, as it assumes large ReMap tables (we refer to this as MemPod_no_AR). For all OTF schemes presented in this section, we used a fixed hotness threshold of 128. We experimented with different thresholds (e.g., 32, 64, and 128) for 4-KB pages and settled on 128 since this threshold provided the best trade-off between the number of pages migrated and the cost of migrations. For both of our OTF migration experiments, one that performs AR using the OS (OTF_OS_AR) and the other that uses our MigC hardware (OTF_HW_AR), we use a 1024 entry ReMap table and start the AR process whenever the table is 50% occupied. We stop migration if the ReMap table does not contain free entries and wait for AR to free up space. More details of the experimental parameters can be found in our previous work [14].

Figure 2 shows the results using static hotness thresholds for all of our workloads[4] for different page migration and AR techniques. A positive *y*-axis value shows performance improvement in terms of instructions per cycle (IPC) as a percentage when compared to a baseline without any page migrations. Likewise, a negative *y*-axis value indicates a performance (or IPC) loss as a percentage compared to the baseline. We separated migration friendly and unfriendly workloads and used different scales for the *y*-axis to make the graphs clearer.

[4]Experimental setup and workloads are described in Section 4.

For page migration friendly workloads, our OTF migration with hardware-based AR technique (OTF_HW_AR) results in 74% IPC improvement on average over the baseline system. It also shows 24% IPC improvement on average over OTF page migration with OS-based AR (OTF_OS_AR).

Our hardware-based migration, OTF_HW_AR, shows higher improvements over other page migration techniques as well: HMA_HS_OS_AR [21] (by 29%) and MemPod_no_AR [27] (by 13%). We included results for migration unfriendly workloads to show that all page migration techniques degrade performance for these workloads, not just our OTF technique, thus justifying our discussion regarding classifying applications as migration friendly and unfriendly in Section 2. It is important to note that for more than half of the migration friendly workloads, even after accounting for all AR overheads (as discussed in Section 2.1), hardware-based AR (OTF_HW_AR) performs better than MemPod_no_AR even when MemPod performs no AR. Next, we compare HMA_HS_OS_AR that used OS-based AR at each epoch with our OTF approach. As shown in Figure 2, our hardware-based AR (OTF_HW_AR) performs better than HMA_HS_OS_AR for all page migration friendly workloads except mix6. In this case, OTF_HW_AR migrated more pages than HMA_HS_OS_AR; the migration benefit of some of the pages is not high. In later sections, we describe adaptive thresholds to monitor the MBQ to control the number of pages migrated.

As expected, within our OTF techniques, hardware-based AR (OTF_HW_AR) performs better than OS-based AR (OTF_OS_AR). The only exception are astar and xalancbmk; the hardware-based AR with smaller overheads is migrating more pages than the OS-based AR methods (since the migrations are paused during AR); however, the additional migrations are not beneficial since these applications are classified as moderately friendly or unfriendly. Epoch-based approaches migrate "hot" pages at the end of an epoch, and some hot pages may have exhausted their usefulness by the time they are migrated. This is one of the reasons for our OTF technique outperforming HMA_HS_OS_AR. However, in epoch-based methods (e.g., [21]), since several pages are migrated at the end of an epoch, AR of all migrated pages can be completed together using the OS, amortizing the cost of AR. Detailed discussions on the performance results can be found in our previous work [14].

2.5 Motivation

The performance results (shown in Figure 2) for different page migration techniques that use static values for hotness thresholds supports our classification of applications as migration friendly and unfriendly shown in Table 1 in Section 2. The very migration friendly applications do show significant performance gains, whereas unfriendly applications show performance losses. With offline analysis such as ours (as described in Section 2 and in our previous work [14]), one could potentially classify an application as either friendly or unfriendly and use page migrations only for migration friendly applications. But some applications exhibit both migration friendly and unfriendly behaviors during different phases of execution. For such applications, relying on offline analysis limits the ability to benefit from page migrations during migration friendly phases of an application.

Consider the behavior of one benchmark, *xalancbmk*, shown in Figure 3. The left side of the figure shows the variation of MBQ and the number of pages migrated over the course of the benchmark execution, using a static "hotness" threshold (in this case 128) in determining when a page is migrated. Both the number of pages migrated and MBQ vary during the execution of the application. Although a large number of pages were migrated during the windows between 70 and 180, the MBQ did not show a significant increase. A properly designed adaptive migration technique can turn on or turn off migration, or adjust the number and frequency of page migrations based on the evolving behavior of an application. The right side of Figure 3 shows the result of using one of our adaptive migration techniques (as described later in Section 3.2). The adaptive

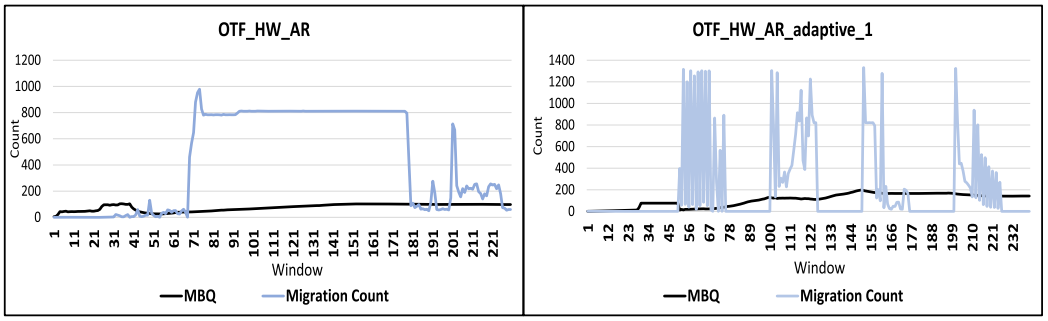

Fig. 3. MBQ and migration count variation of the xalancbmk workload.

technique controls the number of pages migrated based on the MBQ. The figure shows an overall increase in the MBQ (which corresponds in the performance gains).

A second motivation for this study is our desire to minimize overheads due to AR (updating TLBs and PTEs when pages are relocated during migration). Some prior research (e.g., [27]) used very large ReMap tables[5] to eliminate AR. However, the ReMap tables for emerging systems with hundreds of gigabytes to terabytes of memory can become impractical or require additional techniques for the management of large ReMap tables (e.g., caching a small number of ReMap entries while keeping the rest in DRAM). We use small ReMap tables and rely on hardware for AR to minimize OS intervention. As an alternative to using AR, we also explore the idea of reverse migrating previously migrated pages, particularly when they are no longer heavily accessed, making the page migration completely transparent to the OS.

3 ADAPTIVE MIGRATION

The results shown in Figure 2 in Section 2.4 clearly indicate that page migration techniques that use a static or fixed hotness threshold for deciding when a page is considered for migration can lead to performance loss for some applications. In this section, we describe our adaptive migration techniques that adjust the number of pages migrated and even turn off migration by observing the benefits of page migration, eliminating the need for offline analysis for classifying applications as migration friendly and unfriendly. We track the number of pages migrated in a window *twindow* and determine if too many or too few pages are migrated in the window. We also measure the MBQ, which is the average number of accesses to pages recently migrated to faster memories. Using these metrics, we propose two adaptive techniques described in Algorithm 1 and Algorithm 2. Values shown for different variables in these algorithms depend on the page size (4 KB in our experiments) and system parameters such as memory latencies and overheads due to AR. However, they will be constant for a specific system configuration.

3.1 Adaptive Migration Based on Number of Pages Migrated

Algorithm 1 presents an overview of our first adaptive migration technique. We monitor the page migration behaviors over a window, or a threshold_window (i.e. twindow), and dynamically increase or reduce hotness thresholds to be used by our OTF migration techniques. The change is based on how many pages have been migrated during this window. If the count is high (too many pages have been migrated), we double the hotness threshold to reduce future migrations; likewise, if too few pages have been migrated in a *twindow*, we halve the hotness threshold to increase future

[5]Note that ReMap tables keep track of the new locations of migrated pages.

ALGORITHM 1: Adaptive Migration_count technique

```
1:  function ADAPTIVE-MIGRATION(migclk, MBQ, threshold)
2:      migwindow = 100000000;
3:      twindow = 4000000
4:      min_MBQ = 50
5:      max_MBQ = 70
6:      min_mig_count = 160
7:      max_mig_count = 240
8:      min_threshold = 64
9:      max_threshold = 256
10:     if migclk mod twindow == 0 then
11:         if mig_count ≥ max_mig_count && threshold < max_threshold then
12:             threshold *= 2;
13:         else if mig_count ≤ min_min_count && threshold > min_threshold then
14:             threshold /= 2;
15:     if migclk mod migwindow == 0 then
16:         if MBQ ≤ min_MBQ then
17:             pausemigration = true;
18:         else if pausemigration && MBQ ≥ max_MBQ then
19:             pausemigration = false;
20:
```

migrations. In our experiments, we used 4 million cycles as our *twindow*.[6] We also limit the hotness threshold variations between 64 and 256. We increase the threshold if more than 240 pages have been migrated in a window and reduce the threshold if fewer than 160 pages have been migrated in a window. These numbers are based on our observations from our experiments of systems we simulated. These numbers will likely be different for other systems and can be determined from experimental evaluations. We also pause migrations or resume migrations using the MBQ. We define the MBQ as the average number of accesses to pages that were *recently migrated to HBM*. If the MBQ is less than a threshold (min_MBQ), then migrations are halted; migrations are resumed if the MBQ is greater than another threshold (max_MBQ). The decisions regarding pausing and restarting migrations are made only after observing MBQ values over several *twindows*; we refer to this as migration_window (*migwindow*). Our experiments indicated that an average MBQ of less than 90 did not result in performance gains. We use this value to pause migrations. This number (indicating the number of accesses to recently migrated pages in a *twindow*) depends on the overheads to page migrations.

3.2 Adaptive Migration Based on the MBQ

Migration of a page from a slow memory to a faster memory is valuable only if the page receives sufficiently large number of accesses after migration to offset the cost of migration (and AR). Figure 3 in Section 2.5 illustrated that the usefulness of migrated pages (or MBQ) plays a critical role in the overall performance gains achieved. Thus, another approach to dynamically adapt to an application's behavior is to rely on the MBQ as shown in Algorithm 2. Our hardware MigC counts all of the accesses to recently migrated pages in a window (*twindow* or 4 million cycles) and calculates the average MBQ. We decrease (halve) the hotness threshold when the MBQ is high (greater than 130), causing more pages to migrate. We increase (double) the threshold if the MBQ is low (less than 50), to reduce the number of pages migrated. We halt migrations if the MBQ is low

[6]We selected this value to minimize overheads when changing the hotness thresholds. A smaller window results in more rapid adaptation, which in turn causes some overheads in changing the configuration of MigC structures that identify pages ready for migration. Larger windows may be too slow to adapt.

ALGORITHM 2: Adaptive MBQ technique

```
1:  function ADAPTIVE-MIGRATION(migclk, MBQ)
2:      migwindow = 100000000;
3:      twindow = 4000000
4:      min_MBQ = 50
5:      max_MBQ = 70
6:      min_threshold = 64
7:      max_threshold = 256
8:      if migclk mod twindow == 0 then
9:          if MBQ ≤ max_MBQ && threshold < max_threshold then
10:             threshold *= 2;
11:         else if MBQ ≥ min_MBQ && threshold > min_threshold then
12:             threshold /= 2;
13:     if migclk mod migwindow == 0 then
14:         if MBQ ≤ min_MBQ then
15:             pausemigration = true;
16:         else if pausemigration && MBQ ≥ max_MBQ then
17:             pausemigration = false;
18:
```

for several consecutive windows (more than 25 windows). We remind the reader that these specific numbers are based on our experiments and the systems on which we conducted our experiments, and the numbers may be different for different systems.

3.3 Reverse Migration

In most page migration techniques, either epoch-based or OTF migrations, the OS is eventually notified of the migration to modify PTEs and TLBs to reflect new PAs of the migrated pages (since PAs are based on their location). We referred to this process as AR. As described in Section 2.3, and reported in our previous work [14], AR can be very expensive, particularly if performed in software by the OS. The AR can be completely eliminated by using very large ReMap tables, as done in the work of Prodromou et al. [27]. However, the ReMap tables for emerging systems with hundreds of gigabytes to terabytes of memory can become impractical or require additional techniques for the management of large ReMap tables. Instead of using a very large ReMap table, we propose to use small ReMap tables but "reverse migrate" pages back to their original locations, thus making room for more recent hot pages. It should be noted that reverse migration may involve "pairwise" migration if originally a page from faster member was moved to slower memory to make room for a hot page from slower memory.

The reverse migration process is shown in Figure 4. The top left quadrant of the figure shows the identification of hot pages and the top right quadrant shows the original migration, where a PCM page with a PA of 8192 is swapped with a HBM page with a PA of 0. The ReMap table maintains entries for this pair of pages so that CPU requests are correctly directed to their new locations. A previously migrated PCM page to HBM with PA 8192 has turned cold in the bottom left quadrant, and the page is reverse migrated as shown in the bottom right quadrant, where the HBM page that contains the page with the PA of 8192 is swapped with the PCM page with the PA of 0. We rely on MigC hardware for both forward and reverse migrations. The ReMap table entries for these pages will be deleted upon completion of the process (as shown in the bottom right quadrant).

Pages selected for reverse migration can be based several different criteria, such as LRU, or the MBQ (pages with a smaller MBQ are reverse migrated), and the reverse migration can take place on demand or when the ReMap table fills up. Elimination of AR simplifies the complexity of our

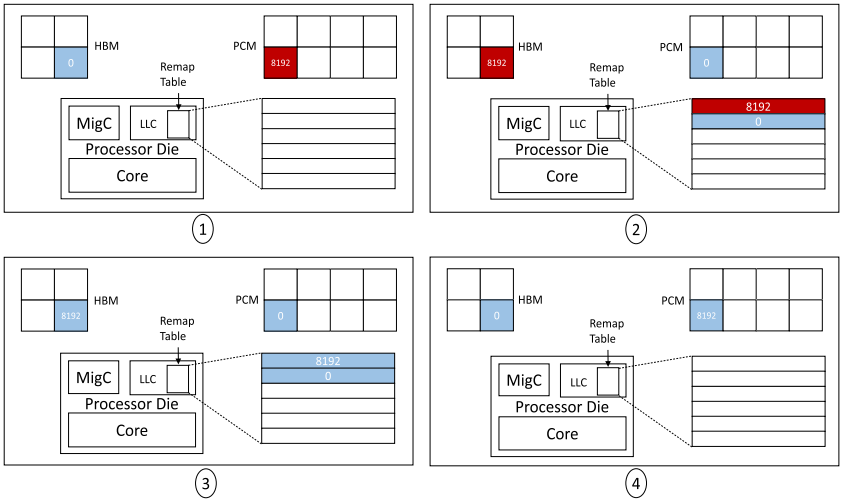

Fig. 4. Reverse migration model.

MigC hardware, since the hardware no longer needs to initiate cache invalidation, shootdown of TLBs, and lock user processes during the reconciliation.

4 EXPERIMENTAL SETUP

In this section, we will describe our simulation setup and the benchmarks we used. This setup and the workloads are the same that we used in our previous work [14] and the same for the experimental results shown previously in Figure 2 in Section 2.4.

4.1 Simulation Infrastructure

We model a 16-core system with a flat-address heterogeneous memory consisting of 1 GB of HBM and 16 GB of PCM using Ramulator [17]. Ramulator is a trace-driven, cycle-level memory simulator with support for a simple multi-core CPU model with cache hierarchies. Each core is four-wide out-of-order issue with 128 reorder buffer entries and operates at 3.2 GHz. The cores have private L1-D caches (32 KB, 4-way, 2-cycles) and shared L2 (16 MB, 16-way, 21-cycles) as LLC. Physical address tags, write-back policies are used for all caches and LLC is inclusive. Ramulator does not model L1-I cache and assumes that non-load/store instructions are executed in 1 cycle. The memory system configuration is provided in Table 2; for timing parameters of HBM, we rely on Kim et al. [17] and for PCM timing on Nair et al. [25]. We modified Ramulator to support a flat-address heterogeneous memory system. A basic address mapping function is added to Ramulator to support this model; it allocates pages to frames of different memories in a round-robin fashion (namely, four pages to faster memory, then four pages to slower memory), as long as there are free frames in faster memory. When faster memory capacity is exhausted, only slower memory frames are assigned. This allocation ensures that pages for all applications span both memory devices, thus necessitating page migration considerations in our experiments. We also made sure that the memory footprints of our benchmarks are at least twice as large as the HBM capacity, requiring the use of both HBM and PCM. We incorporate our MigC unit in Ramulator with all necessary details to perform functions as described in previous sections (Section 2.1). MigC also operates at 3.2 GHz. We assume conventional 4-KB pages. The ReMap table is implemented as a 1024-entry fully associative table. We conservatively assumed an access latency of 10 CPU cycles for the ReMap table

Table 2. Baseline Configuration

Parameter	HBM	PCM
Channels, capacity	8, 1 GB (8 × 128 MB)	2, 16GB (2 × 8 GB)
Memory controller (MC)	1 per channel	1 per channel
Row buffer	2 KB	2 KB
Queue size/MC	RD 32, WR 32, Mig. 32 entries	RD 64, WR 256, Mig. 32 entries
Latency	tCAS-tRCD -tRP-tRAS: 14 ns–14 ns −14 ns–34 ns	Read 80 ns (7.5-ns tPRE + 62.5-ns tSENSE + 10-ns tBUS) Write 250-ns tCWL
Bus/channel	128 bit, 1 GHz	64 bit, 400 MHz

Table 3. Timing Parameters at 3.2-GHz Clock

Task	Time Requirement
ReMap table lookup	10 cycles (after LLC)
Light-weight TLB invalidation at core	300 cycles (round-trip latency to off-chip memory [36])
Page walk	150 cycles
OS reverse mapping	4,480 cycles (measured using FRrace [4] on a real machine running Linux)

(in some experiments, we tested with larger ReMap tables and adjusted access times appropriately). We included all timing overheads (listed in Table 3 assuming 3.2-GHz clock rate) for performing different HW/OS tasks described in Section 2.2. Finally, we used CACTI [24] to model the power rating of the MigC components.

4.2 Workloads

We use 16 multi-programmed SPEC CPU2006 [8] workloads, four multi-threaded benchmarks from the US Department of Energy provided ECP Proxy Applications [10], and the BFS from Graph500 suite [15]. We selected SPEC benchmarks with large memory footprints, at least twice the capacity of HBM. SPEC benchmarks allow us to compare our work with other studies. We profile benchmarks using the PinPlay kit [11] to collect a representative slice of 500M instructions from each of the applications. To make a multi-programmed workload, we run a 16-core Ramulator simulation where each core runs one of the SPEC traces to completion. We either run 16 copies of the same benchmark on 16 cores (each such workload is labeled by the benchmark name in our graphs) or run a random mix of benchmarks on 16 cores (these workloads are labeled as mix1 to mix6 and described in Table 4). The publicly released multi-threaded HPC proxy benchmarks by the US Department of Energy that we used are XSBench [1], LULESH [13], CoMD [22], and miniFE [12]. We ran each HPC benchmark in a 16-thread setup and collected 500M instruction traces for each of the threads using Pin tools [26]. By running traces of the 16 threads of a HPC

Table 4. SPEC Multi-Programmed Mix Workloads

	mix1	mix2	mix3	mix4	mix5	mix6
astar	2x		1x			1x
bzip2		1x	1x	2x		
cactus		2x	2x	1x		
dealII		3x	1x	1x		
gcc	1x		2x	1x		3x
gems		2x	2x	1x		
lbm	2x	3x		1x	6x	1x
leslie			2x	1x		
libq	2x		1x	3x		4x
mcf	3x	2x		1x	5x	
milc	2x		2x	1x		2x
omntpp	1x					3x
soplex	2x	3x		3x	5x	
sphinx	1x		2x			3x

benchmark in Ramulator, we obtain a multi-threaded workload (each such workload is labeled with the name of the benchmark). The memory footprint of the workloads range between 2 and 11 GB, ensuring that the workloads fit in physical memory and do not require access to secondary storage. They are large enough to cause migration but are not unrealistically small.

5 RESULTS AND ANALYSES

In this section, we present the experimental results for our adaptive migration and reverse migration techniques.[7]

5.1 Adaptive Hotness Thresholds

As discussed in Section 2.4, techniques that use fixed hotness thresholds do not perform well on some applications. In this section, we will evaluate our adaptive page migration techniques described in Section 3. As described in that section, we monitor the number of pages migrated in a window: page migration overheads can defeat benefits of migrations if too many pages are migrated. We also monitor the benefits of page migration (or MBQ) by tracking the average number of memory accesses to migrated pages in a window. A low MBQ may indicate that the migrated pages are not heavily accessed and thus the migration was not beneficial. This may be the case for applications that are migration unfriendly. Using these measures, we either change hotness thresholds (to either increase or decrease the number of pages migrated) or completely stop migrations for a duration.

5.2 Evaluation of Adaptive Control Based on Number of Pages Migrated

MigC will monitor the number of pages migrated and sets new threshold values. MigC still performs AR when the ReMap table is more than half full. We compare the IPC using dynamic thresholds with the results obtained using a static threshold of 128, OTF_HW_AR (previously shown in Figure 2). Figure 5 shows the IPC improvements using our first adaptive technique (as described in Algorithm 1) over the baseline with no page migration. The bars labeled OTF_HW_AR_adaptive_1

[7]In Section 2.4, we reported the performance results for migration techniques that use static or fixed hotness thresholds.

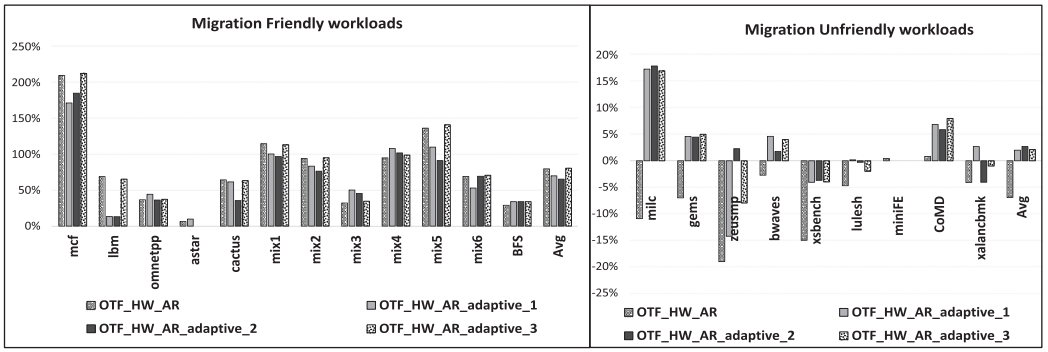

Fig. 5. IPC improvement (%) using adaptive migration policies over a no-migration baseline (negative *y*-axis shows degradation).

refer to the results from our OTF migrations using hardware AR and adapting hotness thresholds based on the number of pages migrated. Our adaptive technique results in 70% performance gains on average over the baseline for migration friendly workloads (the OTF migration with static threshold shows 80% gains over the baseline) but shows on average 2% performance gains over the baseline even for unfriendly workloads (static threshold OTF shows a performance loss of 7%). The adaptive threshold generally reduces the number of pages migrated for migration unfriendly workloads (e.g., milc, gems, bwaves, lulesh, CoMD, xalancmbk) reducing the cost of migrations. However, applications that are friendly (e.g., mcf, lbm, BFS, and some mixed workloads) may see some decrease in performance gains when compared to static threshold–based migrations. This is because when MigC notices that very few pages are migrated in a window, the threshold is reduced to 64 to increase the migrations. However, these additional migrations do not contribute to performance gains. For example, as described in Section 2.5, for mcf, only 3% of pages receive a large number of accesses, and migrating other pages with fewer accesses will not contribute to performance gains but adds to migration and AR costs. It may also be the case that for moderately friendly applications, the adaptive technique may aggressively increase the hotness threshold and reduce the number of pages migrated, even when the application is benefiting from the page migrations. We will explore this later in Section 5.4.

5.3 Evaluation of Adaptive Control Based on the MBQ

Figure 5 includes the results for our second adaptive technique that uses the MBQ to control hotness thresholds, as described in Algorithm 2. The results labeled OTF_HW_AR_adaptive_2 refer to our OTF migrations using hardware for AR and adapting thresholds based on the average access counts to recently migrated pages (i.e., MBQ). On average, our MBQ-based adaptive migration shows a performance gain of 66% over the baseline (compared with 80% gain using static threshold) for migration friendly workloads, and a 3% performance gain on average (compared with a performance loss of 7% with static threshold) for unfriendly workloads. The adaptive technique results in either some performance gains or at least prevent performance losses for migration unfriendly benchmarks: for example, milc, gems, zeusmp, bwaves, CoMD. However, the adaptive MBQ technique results in smaller performance gains for some migration friendly benchmarks when compared to static threshold technique (e.g., mcf, lbm, cactus, and some mixed workloads) for the same reasons outlined in Section 5.2.

To understand the performance of our adaptive algorithms better, we analyzed migration patterns and accesses to migrated pages (Figure 6). We show the average number accesses to pages

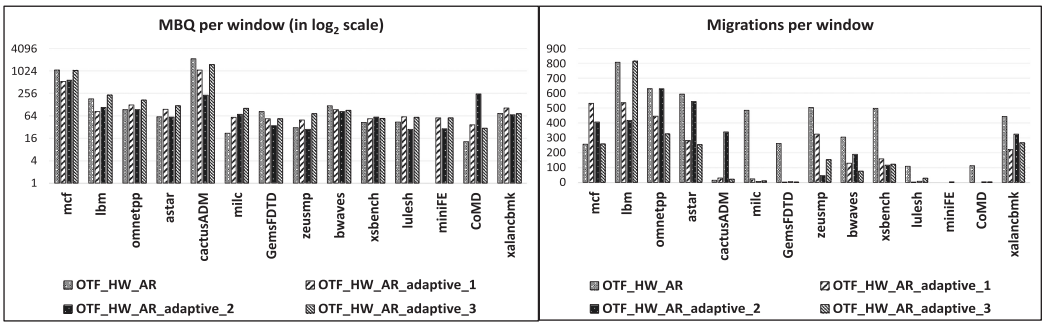

Fig. 6. Traffic data. Left: MBQ per window. Right: Migration count per window.

migrated to HBM in a window of 4 million cycles (or MBQ) on the left and the average number of pages migrated in a window on the right. Please note that the MBQ per window in Figure 6 is shown in $\log_2$ scale. For most of the migration friendly applications, the static threshold technique migrated a very large number of pages, whereas the adaptive techniques aggressively reduced the number of pages migrated and this in turn reduced overall performance gains. Note that our adaptive techniques try to reduce the number of pages migrated to minimize migration overheads. The exception is lbm. As reported in our previous work [14] and listed in Table 1, lbm is actually classified as an only moderately friendly application. For migration unfriendly workloads, adaptive techniques migrated fewer pages and this in turn resulted in a higher average MBQ than static thresholds. The exceptions are gems and bwaves. The number of migrations for gems is greatly reduced in adaptive migration techniques that improved the overall MBQ. For bwaves, the static threshold resulted in a slightly better MBQ than adaptive techniques, but it migrated more than twice as many pages: the migration overheads defeat the MBQ advantage. These results directly translate into performance gains shown in Figure 5.

5.4 Evaluation of Adaptive Migrations Based on a Combined Technique

Based on these observations regarding the two adaptive techniques presented thus far, we explored a third adaptive technique that combines these two methods. This is shown in Algorithm 3. In this method, when the number of pages migrated exceeds the threshold, but the MBQ is high, we will not increase the hotness threshold (which in turn reduces the number of pages migrated), since the high MBQ indicates that page migration is beneficial even when a large number of pages is migrated.

The results of this combined adaptive technique are also included in Figure 5, which are labeled as OTF_HW_AR_adaptive_3. The figure shows that the combined approach achieves almost the same level of performance as the static threshold technique for (very) migration friendly applications. However, the combined adaptive technique outperforms static threshold and the other adaptive migration techniques for moderately friendly and unfriendly workloads. On average, our combined adaptive migration shows a performance gain of 81% over the baseline (compared with 80% gain using a static threshold) for migration friendly workloads, and a 2% performance gain on average (compared with a performance loss of 7% with a static threshold) for unfriendly workloads.

Our adaptive techniques achieve these goals without having to perform offline analyses. Our adaptive techniques are very simple to implement. We either count the number of pages migrated in a given time window or measure the number of accesses to pages recently migrated to HBM. It may be possible to investigate more intelligent adaptive techniques, but they will likely be complex to implement.

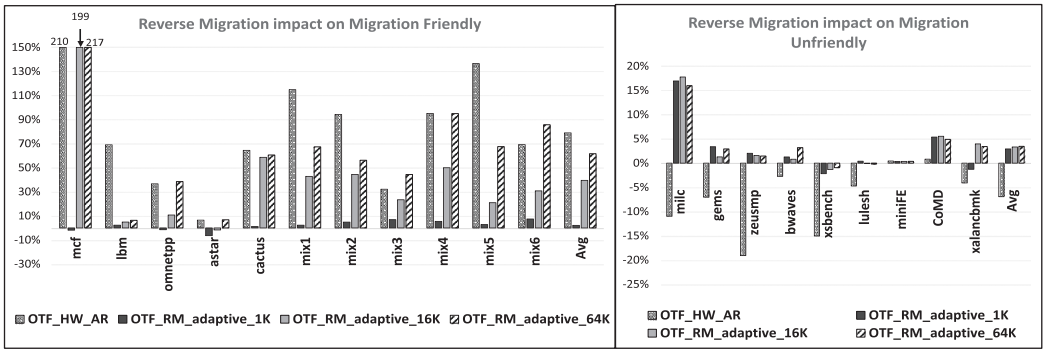

Fig. 7. IPC improvement (%) of reverse migration over a no-migration baseline with different ReMap table sizes (negative *y*-axis shows degradation).

ALGORITHM 3: Combined Adaptive technique (initialized values depend on the page size)

```
1:  function ADAPTIVE-MIGRATION(migclk, MBQ)
2:      migwindow = 100000000;
3:      twindow = 4000000
4:      min_MBQ = 50
5:      max_MBQ = 70
6:      upper_MBQ = 100
7:      min_threshold = 32
8:      max_threshold = 256
9:      if migclk mod twindow == 0 then
10:         if (mig_count ≥ max_mig_count) && (threshold < max_threshold) && (MBQ ≤
    upper_MBQ) then
11:             threshold *= 2;
12:         else if (mig_count ≤ min_mig_count) && (threshold > min_threshold) && (MBQ ≤
    max_MBQ) then
13:             threshold /= 2;
14:     if migclk mod migwindow == 0 then
15:         if MBQ ≤ min_MBQ then
16:             pausemigration = true;
17:         else if pausemigration && MBQ ≥ max_MBQ then
18:             pausemigration = false;
19:
```

5.5 Reverse Migration of Pages

Next, we explored reverse migration of pages to eliminate AR. As the ReMap table fills up, instead of reconciling addresses of pages and removing ReMap entries, we migrate pages back to their original locations and remove the corresponding ReMap entries, as described in Section 3.3. For our experiments, we select pages based on the MBQ: pages with the least number of accesses in a window are candidates for reverse migration, after making sure that migrated pages remain in HBM for a minimum duration (in our case 1 million cycles) to avoid reverse migrating pages too early. If no pages are suitable for reverse migration, page migration is temporarily delayed. Figure 7 shows the results from our reverse migration experiments.

We still use adaptive thresholds relying on the MBQ as described in Algorithm 2. For comparison purposes, Figure 7 includes data for migration based on a static threshold of 128 (previously shown in Figure 2: these results are labeled as OTF_HW_AR (with 1024 ReMap entries). For reverse migration experiments, we vary the ReMap table sizes between 1K and 64K entries. The

results are labeled with ReMap table sizes. For example, OTF_RM_adaptive_1K refers to OTF migrations with reverse migrations, using adaptive MBQ-based thresholds and a ReMap table with 1024 entries. For migration friendly workloads, on average the reverse migration technique shows performance gains of 2%, 41%, and 57% with 1K, 16K, and 64K ReMap tables, respectively, over the baseline (compared with 74% gains using a static threshold). A smaller ReMap table causes frequent migrations and reverse migrations leading to excessive data movement. A larger ReMap table results in performance gains for most applications when compared to the baseline with no page migrations. For some migration friendly benchmarks (e.g., lbm), reverse migration may cause heavily accessed pages to be migrated and reverse migrated several times. AR eliminates such repeated migrations since heavily accessed pages are likely to be permanently moved to HBM (and PAs reconciled). For most workloads, larger ReMap tables provide sufficient space for heavily accessed pages to remain in the ReMap table for longer periods of time, requiring fewer reverse migrations. For example, consider mcf; based on our characterization [14], only 3% of the pages account for 90% of the memory accesses. So, having such large ReMap table sizes helped in reducing reverse migrations. Except for some very friendly workloads like mix1, mix2, mix5, and lbm, reverse migration (OTF_RM_adaptive_64K) performs on par or even better than OTF_HW_AR. Even for migration unfriendly workloads, reverse migration shows performance gains of about 3% for all ReMap table sizes (compared with 7% performance loss using static threshold). Our adaptive techniques either limit or stop page migrations for these migration unfriendly applications. Thus, the size of the ReMap table has very little impact on the performance of migration unfriendly workloads. We believe that reverse migrations can be a viable option to HMA systems, particularly when the AR is costly. Larger ReMap tables are justified since reverse mapping eliminates the complexity and overheads of AR. A 64K ReMap table would require about 1.3 MB of storage. This table can be placed in HBM while caching a small portion inside MigC. This is similar to other reported studies [6, 27, 33]; in those studies, the ReMap tables were much larger than what we are proposing.

5.6 Subpage Migration

Although not actually implemented in this study, reverse migration may be useful in systems with very large pages. As the physical memory sizes increase, traditional 4K byte pages also increase the sizes of page tables and TLBs. Page table and TLB sizes can be reduced by using larger pages, say 64-KB, 1-MB, 2-MB, or even 1-GB pages.

However, it should be noted that only small portions of a large (or huge) page is likely to be hot, whereas other portions are not heavily accessed. It will be wasteful to migrate the entire page in such cases. So, instead of migrating large pages, one may consider migrating only hot subpages of large pages. Such subpage migrations require tracking the physical location of subpages. ReMap tables can be extended for this purposes: each entry in the ReMap table can contain a bit map to indicate if a subpage is migrated or not. But AR becomes very cumbersome; we need to migrate rest of the subpages before updating PTE and TLB entries. Instead, the use of reverse migration can provide an alternative to AR, making subpage migration a potentially viable method. Migration (and reverse migration) can take place at smaller subpage granularity (1 KB, 2 KB), as these result in better performance for some benchmarks. We will explore reverse migration of subpages for systems using huge pages in the future.

5.7 Energy Consumption

Figure 8 shows the dynamic energy savings that are achieved using adaptive and reverse migration techniques when compared with the baseline (with no page migrations). For the baseline system, we measure energy for all demand requests to faster and slower memory. Energy consumption for our techniques account for energy for demand requests, energy consumed for

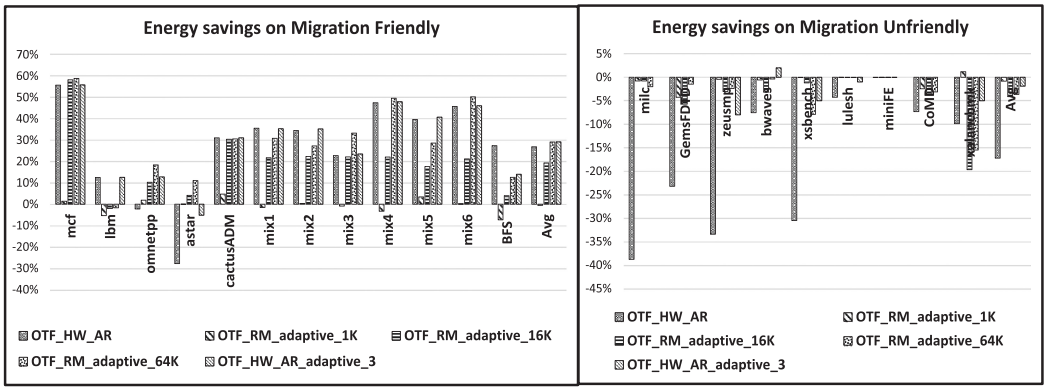

Fig. 8. Energy savings (%) of adaptive and reverse migration over a no-migration baseline (negative *y*-axis shows degradation).

Table 5. Memory Energy Parameters

Memory	Access Energy
HBM	3.92 pj/bit
PCM	Read 42 pj/bit Write 140 pj/bit

migration of pages (which requires additional accesses to memory), and energy consumed by the MigC hardware to manage migrations and AR. It should be noted that the HBM energy consumption is much lower than PCM, as shown in Table 5. As the number of accesses to the HBM increases with the decrease in the accesses to PCM when pages are migrated (even accounting for the migration cost itself), noticeable energy savings can be observed. All of our OTF migration techniques show energy savings for migration friendly workloads: the hardware AR technique (OTF_HW_AR) consumes on average 27% less energy, whereas the adaptive techniques (OTF_HW_AR_adaptive_1, OTF_HW_AR_adaptive_2 and OTF_HW_AR_adaptive_3) consume 25%, 22%, and 29% less energy than the baseline, respectively. As shown in Section 5.1, the execution performance of adaptive techniques (OTF_HW_AR_adaptive_1, OTF_HW_AR_adaptive_2) is slightly lower than OTF_HW_AR, which uses a static threshold. This in turn also results in slightly higher energy consumption for adaptive techniques, whereas OTF_HW_AR_adaptive_3 performs slightly better than OTF_HW_AR and consumes 2% less energy than OTF_HW_AR. The reverse migration techniques with 1K, 16K, and 64K ReMap table sizes (OTF_RM_adaptive_1K, OTF_RM_adaptive_16K, OTF_RM_adaptive_64K) show 0%, 21%, and 29% energy savings, respectively, over the baseline with no page migrations. As stated, OTF_RM_adaptive_1K causes frequent migrations and reverse migrations even for heavily accessed pages that otherwise would have been reconciled. Thus, the benefits of the migrations are negated by the reverse migrations. For migration unfriendly workloads, our adaptive techniques save power up to 15% compared to the OTF_HW_AR technique on average. In the case of the reverse migration technique with 16K and 64K ReMap tables (OTF_RM_adaptive_16K and OTF_RM_adaptive_64K), ReMap table sizes are large enough to hold hot pages for longer periods, minimizing the need for reverse migration. This in turn results in improved energy savings. In some cases, the benefit of large ReMap table size is sufficient to show energy savings that equal the improvements achieved with the best

adaptive technique (OTF_HW_AR_adaptive_3). In our energy computations, we have accounted for the additional power needed for large ReMap table sizes.

6 RELATED WORK

There have been many studies on page migration techniques for flat-address heterogeneous memory systems (HMA). They propose different approaches to solve the general challenges associated with page migration, namely selecting candidate pages to migrate, determining migration frequency, and managing migration metadata. Here we will review only the works that are most closely related to our research. Meswani et al. [21] presented a study where page migration in HMA is accomplished by a hardware/software (we refer to it as HMA-HS) mixed approach. The hardware keeps track of the page access counts over a fixed-length epoch, and at the end of each epoch, the hottest pages residing in slow memory are migrated to fast memory by the OS, updating PAs of the migrating pages. There is no restriction on where (i.e., which HBM page frame) a hot page can be migrated. Since OS-based address update incurs large overheads, the authors chose a longer epoch to reduce frequent OS interventions. However, it has been observed that page migration at shorter intervals is more beneficial than waiting for longer epoch times [27, 33], since migrating hot pages sooner results in more beneficial accesses to the migrated pages.

AR can be avoided using a very large ReMap table that contains the new locations of migrated pages. The size and management of this ReMap table presents a new challenge. A number of different approaches have been proposed to keep this table in memory while using a small on-chip cache for recently accessed entries [6, 18, 27, 33]. Sim et al. [33] used a transparent hardware-based management of flat-address memory, which restricts where a migrated page can be placed to reduce the ReMap table: a set of slow memory pages compete for a single fast memory page. This can reduce potential benefits since only one of the slow memory pages from a given set can be migrated to fast memory even if all of them are heavily accessed. Chou et al. [6] proposed a somewhat similar idea of intra-set migration named *CAMEO*, where the migration is done at finer granularity (cache line size) and the migration candidate is chosen on each slow memory access, whereas Chameleon [19], depending on the application requirement, dynamically reconfigures the parts of stacked DRAM as flat-address memory or cache. Prodomou et al. [27] proposed MemPod, which provides more flexibility on page relocation than in other works [6, 33]. On-chip MCs for fast and slow memories are grouped into "Pods," and only intra-Pod epoch-based page migration is allowed. A low-cost counter is used to keep track of recently accessed hot pages. A more recent work, PageSeer [18] proposes extensions to MCs that initiate page swaps between slow and fast memories based on TLB misses. To use the 3D-DRAM capacity efficiently, Ryoo et al. [32] proposed a sub-block-based page migration and co-location of sub-blocks of two different pages in an interleaved fashion. However, this scheme only allows migration of sub-blocks within same congruence group and requires large SRAM tables to keep track of sub-blocks with bit vectors. In the work of Ryoo et al. [31], the granularity of the data migrated depends on the contiguity of accesses in the VA space. In our work, migration granularity is fixed; however, we have included an evaluation of how the page size impacts the performance of page migrations.

In our proposal, we migrate a page immediately when it receives a sufficient number of memory accesses, unlike any epoch-based schemes described earlier. We allow full flexibility in page relocation like HMA-HS [21] and keep a small on-chip ReMap table for address redirection. A similar approach has been proposed by Ramos et al. [29], which also performs an OTF type of migration with periodical reconciliation of remapping table entries and OS memory mappings. However, in the work of Ramos et al. [29], the migration and reconciliation processes are separate phases since the reconciliation is completely handled by the OS. In our proposal, we perform AR with help of specialized hardware, and hence these processes can progress concurrently. Furthermore, our

migration candidate choice scheme is simpler than the multi-queue scheme used by Ramos et al. [29]. We also study dynamic adjustments to page migrations: we change hotness thresholds to reduce or increase the number of pages migrated or pause migrations when insignificant benefits are observed. We also explore reverse migration of pages to eliminate AR.

Yu et al. [37] propose a technique (called *Banshee*) for transferring pages between off-chip memory and on-chip DRAM cache. Instead of storing tags for DRAM cache, Banshee uses Tag-buffers, somewhat similar to our ReMap tables. As the Tag-buffer fill up, the user programs are stopped and the OS updates the TLB (and PTE) entries to reflect the new location of pages—again somewhat similar to our AR. However, unlike our hardware orchestrated approach, they rely on the OS for AR. Moreover, Banshee assumes that the two levels of memory (on-chip and off-chip DRAMs) have similar latencies but differ in bandwidth, whereas we assume memory technologies with significantly different latencies and bandwidths.

Conventional non-uniform memory access (NUMA) systems with multi-socket CPU and homogeneous memory emulate data migration between local and remote nodes. Accesses to data within local memory are faster compared to remote memory locations. The main difference between NUMA and HMA data migration is that NUMA migration occurs when cores from different sockets need to work on the same data. This issue of on-demand migration is mitigated in other works [34, 38] by running multiple threads that share data on the same node, whereas in HMA, pages are migrated in a single-node system.

7 CONCLUSION AND FUTURE WORK

In this work, we extended our previous study [14] of an OTF page migration technique that migrates a page from slow NVM to fast memory (e.g., HBM) as soon as the page becomes "hot." For this contribution, we employ adaptive migration techniques that dynamically change the frequency and the number of pages migrated, and pause or resume migrations based on the memory access behavior of applications. We use three different adaptive techniques. First, we monitor the number of pages migrated in a given observation window (mig_count_adaptive). If the number exceeds a threshold, we increase the hotness threshold so that the number of migrations are reduced (likewise, we reduce the threshold if the number of pages migrated is below a threshold). Second, we monitor the MBQ (the average number of accesses to page after migration, which indicates the usefulness of a migration, MBQ_adaptive) and either increase or reduce migrations based on the MBQ. We pause migrations temporarily if the MBQ is too low and resume migrations if the MBQ increases. Third, we explore a combination of the previous two approaches: we monitor the number of pages migrated in a window and increase the hotness threshold if too many pages are migrated and the MBQ is low. Likewise, we reduce the threshold when the number of pages migrated is low and the MBQ is high. The first technique has resulted in an average of 71% IPC improvement over the baseline for migration friendly workloads, but more importantly, our technique has shown on average 2% performance gains over the baseline even for unfriendly workloads. In terms of energy consumption, this technique has achieved 25% energy savings compared to the baseline for migration friendly workloads but has consumed 3% more energy than the baseline for unfriendly workloads. This should be compared with the static threshold OTF migration technique, which consumes 17% more energy than the baseline. The second technique that monitors the MBQ has achieved an average performance gain of 65% over the baseline for migration friendly workloads, 3% gains for migration unfriendly workloads, and 22% energy savings for migration friendly workloads while consuming 2% more energy for migration unfriendly applications over the baseline. The third technique that combines the previous two adaptive techniques resulted in 81% performance gains on average over the baseline for migration friendly workloads and 2% gains even for migration unfriendly workloads. Our adaptive techniques eliminates the need for offline

profiling of applications to categorize them as a migration friendly or unfriendly. We believe that additional adaptive techniques may further improve the performance gains with page migrations in multi-level heterogeneous memories.

We also experimented with reverse migration of pages, eliminating AR (making the page migration completely transparent to the OS). When the ReMap table is almost full, ReMap entries are freed for new page migrations by reverse migrating pages with a low MBQ back to their original locations. We tested with ReMap table sizes of 1K, 16K, and 64K entries. The reverse migration technique has achieved performance improvements of 2%, 41%, and 57%, respectively, for migration friendly workloads and 3%, 3%, and 3%, respectively, for migration unfriendly workloads. Regarding energy savings, reverse migration technique has achieved 0%, 21%, and 29% for migration friendly workloads and has consumed 1%, 4%, and 4% for migration unfriendly workloads. It can be observed that the static OTF technique (OTF_HW_AR) achieved 28% energy savings, which is lower than the reverse migration technique with 64K ReMap table entries by 1% but requires more area for the extra ReMap table. We believe that reverse migration can be used when dealing with very large pages. Since migrating huge pages can be very expensive, one can consider migrating only hot subpages of a huge page and reverse migrate the subpages as they become cold. We plan to explore such subpage migrations in the future.

ACKNOWLEDGMENTS

The authors would like to acknowledge Nuwan Jayasena and Mike Ignatowski of AMD for their feedback and suggestions throughout the conduction of this research.

REFERENCES

[1] John R. Tramm, Andrew R. Siegel, Tanzima Islam, and Martin Schulz. 2014. XSBench–The development and verification of a performance abstraction for Monte Carlo reactor analysis. In *Proceedings of the Role of Reactor Physics toward a Sustainable Future (PHYSOR'14)*. https://www.mcs.anl.gov/papers/P5064-0114.pdf.

[2] Amro Awad, Arkaprava Basu, Sergey Blagodurov, Yan Solihin, and Gabriel H. Loh. 2017. Avoiding TLB shootdowns through self-invalidating TLB entries. In *Proceedings of the 2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT'17)*. IEEE, Los Alamitos, CA, 273–287.

[3] Chaim Bendelac and Panos Kokkalis. 2017. SAP HANA Memory Usage Explained. Retrieved January 20, 2019 from https://www.sap.com/documents/2016/08/205c8299-867c-0010-82c7-eda71af511fa.html.

[4] Tim Bird. 2009. Measuring function duration with FTrace. In *Proceedings of the Linux Symposium*. 47–54.

[5] Daniel P. Bovet and Marco Cesati. 2005. *Understanding the Linux Kernel: From I/O Ports to Process Management*. O'Reilly Media.

[6] Chiachen Chou, Aamer Jaleel, and Moinuddin K. Qureshi. 2014. CAMEO: A two-level memory organization with capacity of main memory and flexibility of hardware-managed cache. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, Los Alamitos, CA, 1–12.

[7] HMC Consortium. 2018. Hybrid Memory Cube Consortium. Retrieved July 27, 2018 from http://hybridmemorycube.org/.

[8] Standard Performance Evaluation Corporation. 2015. SPEC CPU 2006. Retrieved February 9, 2021 from https://www.spec.org/cpu2006/.

[9] Joint Electron Devices Engineering Council. 2018. 3D ICs. Retrieved July 27, 2018 from http://www.jedec.org/category/technology-focus-area/3d-ics-0.

[10] US Department of Energy. 2018. US Department of Energy ECP Proxy Application Suite. Retrieved February 9, 2021 from https://proxyapps.exascaleproject.org/ecp-proxy-apps-suite/.

[11] Harish Patil. 2018. PinPlay. Retrieved February 9, 2021 from https://software.intel.com/en-us/articles/program-recordreplay-toolkit.

[12] M. Heroux and S. Hammond. 2015. MiniFE: finite element solver. https://portal.nersc.gov/project/CAL/designforward.htm#MiniFE.

[13] R. D. Hornung, J. A. Keasler, and M. B. Gokhale. 2011. *Hydrodynamics Challenge Problem*. Technical Report. Lawrence Livermore National Lab, Livermore, CA.

[14] Mahzabeen Islam, Shashank Adavally, Marko Scrbak, and Krishna Kavi. 2020. On-the-fly page migration and address reconciliation for heterogeneous memory systems. *ACM Journal on Emerging Technologies in Computing Systems* 16, 1 (2020), Article 10. http://csrl.cse.unt.edu/kavi/Research/JETC-2019.pdf.

[15] Jewillco. 2015. Graph500-v2-spec. Retrieved February 9, 2021 from https://github.com/graph500/graph500/tree/v2-spec.

[16] Kimberly Keeton. 2017. *Memory-Driven Computing*. USENIX Association, Santa Clara, CA.

[17] Y. Kim, W. Yang, and O. Mutlu. 2016. Ramulator: A fast and extensible DRAM simulator. *IEEE Computer Architecture Letters* 15, 1 (2016), 45–49.

[18] A. Kokolis, D. Skarlatos, and J. Torrellas. 2019. PageSeer: Using page walks to trigger page swapps in hybrid memory systems. In *Proceedings of the 25th IEEE International Symposium on High Performance Computer Architecture*. IEEE, Los Alamitos, CA.

[19] J. B. Kotra, H. Zhang, A. R. Alameldeen, C. Wilkerson, and M. T. Kandemir. 2018. CHAMELEON: A dynamically reconfigurable heterogeneous memory system. In *Proceedings of the 2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'18)*. 533–545. DOI: https://doi.org/10.1109/MICRO.2018.00050

[20] E. Kultursay, M. Kandemir, A. Sivasubramaniam, and O. Mutlu. 2013. Evaluating STT-RAM as an energy-efficient main memory alternative. In *Proceedings of the 2013 IEEE International Symposium on Performance Analysis of Systems Software*. IEEE, Los Alamitos, CA.

[21] Mitesh R. Meswani, Sergey Blagodurov, David Roberts, John Slice, Mike Ignatowski, and Gabriel H. Loh. 2015. Heterogeneous memory architectures: A HW/SW approach for mixing die-stacked and off-package memories. In *Proceedings of the 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA'15)*. IEEE, Los Alamitos, CA, 126–136.

[22] Jamaludin Mohd-Yusof, Sriram Swaminarayan, and Timothy C. Germann. 2013. Co-design for molecular dynamics: An exascale proxy application. https://www.lanl.gov/orgs/adtsc/publications/science_highlights_2013/docs/Pg88_89.pdf.

[23] David Mosberger and Stephane Eranian. 2001. *IA-64 Linux Kernel: Design and Implementation*. Prentice Hall PTR.

[24] N. Muralimanohar, Rajeev Balasubramonian, and N. Jouppi. 2007. CACTI 6 . 0 : A tool to understand large caches. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.3834&rep=rep1&type=pdf.

[25] Prashant J. Nair, Chiachen Chou, Bipin Rajendran, and Moinuddin K. Qureshi. 2015. Reducing read latency of phase change memory via early read and Turbo Read. In *Proceedings of the 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA'15)*. IEEE, Los Alamitos, CA, 309–319.

[26] Osnat Levi (Intel). 2018. Pin - A Dynamic Binary Instrumentation Tool. Retrieved February 9, 2021 from https://software.intel.com/en-us/articles/pin-a-dynamic-binary-instrumentation-tool.

[27] Andreas Prodromou, Mitesh Meswani, Nuwan Jayasena, Gabriel Loh, and Dean M. Tullsen. 2017. MemPod: A clustered architecture for efficient and scalable migration in flat address space multi-level memories. In *Proceedings of the 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA'17)*. IEEE, Los Alamitos, CA, 433–444.

[28] Moinuddin K. Qureshi, Sudhanva Gurumurthi, and Bipin Rajendran. 2011. Phase change memory: From devices to systems. *Synthesis Lectures on Computer Architecture* 6, 4 (2011), 1–134.

[29] Luiz E. Ramos, Eugene Gorbatov, and Ricardo Bianchini. 2011. Page placement in hybrid memory systems. In *Proceedings of the International Conference on Supercomputing*. ACM, New York, NY, 85–95.

[30] Bogdan F. Romanescu, Alvin R. Lebeck, Daniel J. Sorin, and Anne Bracy. 2010. UNified instruction/translation/data (UNITD) coherence: One protocol to rule them all. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA'10)*.

[31] Jee Ho Ryoo, Lizy K. John, and Arkaprava Basu. 2018. A case for granularity aware page migration. In *Proceedings of the 32nd International Conference on Supercomputing (ICS'18)*. 352–362. DOI: https://doi.org/10.1145/3205289.3208064

[32] Jee Ho Ryoo, Mitesh R. Meswani, Andreas Prodromou, and Lizy K. John. 2017. SILC-FM: Subblocked interleaved cache-like flat memory organization. In *Proceedings of the 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA'17)*. IEEE, Los Alamitos, CA, 349–360.

[33] Jaewoong Sim, Alaa R. Alameldeen, Zeshan Chishti, Chris Wilkerson, and Hyesoon Kim. 2014. Transparent hardware management of stacked dram as part of memory. In *Proceedings of the 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'14)*. IEEE, Los Alamitos, CA, 13–24.

[34] F. Song, S. Moore, and J. Dongarra. 2009. Analytical modeling and optimization for affinity based thread scheduling on multicore systems. In *Proceedings of the 2009 IEEE International Conference on Cluster Computing and Workshops*. 1–10. DOI: https://doi.org/10.1109/CLUSTR.2009.5289173

[35] ChunYi Su, David Roberts, Edgar A. León, Kirk W. Cameron, Bronis R. de Supinski, Gabriel H. Loh, and Dimitrios S. Nikolopoulos. 2015. HpMC: An energy-aware management system of multi-level memory architectures. In *Proceedings of the 2015 International Symposium on Memory Systems*. ACM, New York, NY, 167–178.

[36] Carlos Villavieja, Vasileios Karakostas, Lluis Vilanova, Yoav Etsion, Alex Ramirez, Avi Mendelson, Nacho Navarro, Adrian Cristal, and Osman S. Unsal. 2011. Didi: Mitigating the performance impact of TLB shootdowns using a shared TLB directory. In *Proceedings of the 2011 International Conference on Parallel Architectures and Compilation Techniques (PACT'11)*. IEEE, Los Alamitos, CA, 340–349.

[37] Xiangyao Yu, Christopher J. Hughes, Nadathur Satish, Onur Mutlu, and Srinivas Devadas. 2017. Banshee: Bandwidth-efficient DRAM caching via software/hardware cooperation. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'17)*. ACM, New York, NY, 1–14. DOI: https://doi.org/10.1145/3123939.3124555

[38] I. Ştirb. 2018. NUMA-BTLP: A static algorithm for thread classification. In *Proceedings of the 2018 5th International Conference on Control, Decision, and Information Technologies (CoDIT'18)*. 882–887. DOI: https://doi.org/10.1109/CoDIT.2018.8394925

Received December 2019; revised August 2020; accepted December 2020