

Received July 12, 2021, accepted August 31, 2021, date of publication September 8, 2021, date of current version October 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110993

DAMOV: A New Methodology and Benchmark Suite for Evaluating Data Movement Bottlenecks

GERALDO F. OLIVEIRA¹, JUAN GÓMEZ-LUNA¹, (Member, IEEE), LOIS OROSA¹, (Member, IEEE), SAUGATA GHOSE², (Member, IEEE), NANDITA VIJAYKUMAR³, IVAN FERNANDEZ^{1,4}, MOHAMMAD SADROSADATI¹, AND ONUR MUTLU¹, (Fellow, IEEE)

¹Department of Information Technology and Electrical Engineering (D-ITET), ETH Zürich, 8092 Zürich, Switzerland

²Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL 61801, USA

³Department of Computer Science, University of Toronto, Toronto, ON M5S 2B1, Canada

⁴Department of Computer Architecture, University of Malaga, 29016 Málaga, Spain

Corresponding author: Geraldo F. Oliveira (geraldod@inf.ethz.ch)

This work was supported in part by the SAFARI Research Group's industrial partners through ASML, Facebook, Google, Huawei, Intel, Microsoft, VMware, and Semiconductor Research Corporation.

ABSTRACT Data movement between the CPU and main memory is a first-order obstacle against improving performance, scalability, and energy efficiency in modern systems. Computer systems employ a range of techniques to reduce overheads tied to data movement, spanning from traditional mechanisms (e.g., deep multi-level cache hierarchies, aggressive hardware prefetchers) to emerging techniques such as Near-Data Processing (NDP), where some computation is moved close to memory. Prior NDP works investigate the root causes of data movement bottlenecks using different profiling methodologies and tools. However, there is still a lack of understanding about the key metrics that can identify different data movement bottlenecks and their relation to traditional and emerging data movement mitigation mechanisms. Our goal is to methodically identify potential sources of data movement over a broad set of applications and to comprehensively compare traditional compute-centric data movement mitigation techniques (e.g., caching and prefetching) to more memory-centric techniques (e.g., NDP), thereby developing a rigorous understanding of the best techniques to mitigate each source of data movement. With this goal in mind, we perform the first large-scale characterization of a wide variety of applications, across a wide range of application domains, to identify fundamental program properties that lead to data movement to/from main memory. We develop the first systematic methodology to classify applications based on the sources contributing to data movement bottlenecks. From our large-scale characterization of 77K functions across 345 applications, we select 144 functions to form the first open-source benchmark suite (DAMOV) for main memory data movement studies. We select a diverse range of functions that (1) represent different types of data movement bottlenecks, and (2) come from a wide range of application domains. Using NDP as a case study, we identify new insights about the different data movement bottlenecks and use these insights to determine the most suitable data movement mitigation mechanism for a particular application. We open-source DAMOV and the complete source code for our new characterization methodology at <https://github.com/CMU-SAFARI/DAMOV>.

INDEX TERMS Benchmarking, data movement, energy, memory systems, near-data processing, performance, processing-in-memory, workload characterization, 3D-stacked memory.

I. INTRODUCTION

Today's computing systems require moving data from main memory (consisting of DRAM) to the CPU cores so that computation can take place on the data. Unfortunately, this *data movement* is a major bottleneck for system

The associate editor coordinating the review of this manuscript and approving it for publication was Tae Hyoung Kim.

performance and energy consumption [1], [2]. DRAM technology scaling is failing to keep up with the increasing memory demand from applications [2]–[29], resulting in significant latency and energy costs due to data movement [1]–[3], [5], [6], [30]–[49]. High-performance systems have evolved to include mechanisms that aim to alleviate data movement's impact on system performance and energy consumption, such as deep cache hierarchies and

aggressive prefetchers. However, such mechanisms not only come with significant hardware cost and complexity, but they also often fail to hide the latency and energy costs of accessing DRAM in many modern and emerging applications [1], [5], [50]–[52]. These applications' memory behavior can differ significantly from more traditional applications since modern applications often have lower memory locality, more irregular access patterns, and larger working sets [36], [45], [46], [53]–[61]. One promising technique that aims to alleviate the data movement bottleneck in modern and emerging applications is Near-Data Processing (NDP) [1], [33], [34], [46]–[48], [54], [55], [59]–[118],¹ where the cost of data movement to/from main memory is reduced by placing computation capability close to memory. In NDP, the computational logic close to memory has access to data that resides in main memory with significantly higher memory bandwidth, lower latency, and lower energy consumption than the CPU has in existing systems. There is very high bandwidth available to the cores in the logic layer of 3D-stacked memories, as demonstrated by many past works (e.g., [1], [46], [59], [60], [62]–[64], [67]–[69], [74], [76], [99], [119]). To illustrate this, we use the STREAM Copy [120] workload to measure the peak memory bandwidth the host CPU and an NDP architecture with processing elements in the logic layer of a single 3D-stacked memory (e.g., Hybrid Memory Cube [73]) can leverage.² We observe that the peak memory bandwidth that the NDP logic can leverage (431 GB/s) is $3.7\times$ the peak memory bandwidth that the host CPU can exploit (115 GB/s). This happens since the external memory bandwidth is bounded by the limited number of I/O pins available in the DRAM device [121].

Many recent works explore how NDP can benefit various application domains, such as graph processing [46], [47], [54], [63], [74], [93], [122]–[126], machine learning [1], [61], [69], [70], [84], [85], [103], bioinformatics [59], [60], [68], databases [55], [61], [63], [66], [67], [74], [86], [102], security [71], [105], [106], data manipulation [49], [86], [88], [89], [127]–[130], and mobile workloads [1], [61]. These works demonstrate that simple metrics such as last-level CPU cache Misses per Kilo-Instruction (MPKI) and Arithmetic Intensity (AI) are useful metrics that serve as a proxy for the amount of data movement an application experiences. These metrics can be used as a potential guide for choosing when to apply data movement mitigation mechanisms such as NDP. However, such metrics (and the corresponding insights) are often extracted from a small set of applications, with similar or not-rigorously-analyzed data movement characteristics. Therefore, it is difficult to generalize the metrics and insights these works provide to a broader set of applications, making it unclear what different metrics can reveal about a new (i.e., previously uncharacterized) application's data movement behavior (and how to mitigate its associated data movement costs).

We illustrate this issue by highlighting the limitations of two different methodologies commonly used to identify memory bottlenecks and often used as a guide to justify the use of NDP architectures for an application: (a) analyzing a roofline model [131] of the application, and (b) using last-level CPU cache MPKI as an indicator of NDP suitability of the application. The roofline model correlates the computation requirements of an application with its memory requirements under a given system. The model contains two *roofs*: (1) a diagonal line ($y = \text{Peak Memory Bandwidth} \times \text{Arithmetic Intensity}$) called the *memory roof*, and (2) a horizontal line ($y = \text{Peak System Throughput}$) called the *compute roof* [131]. If an application lies under the memory roof, the application is classified as *memory-bound*; if an application lies under the compute roof, it is classified as *compute-bound*. Many prior works [99], [103], [132]–[144] employ this roofline model to identify memory-bound applications that can benefit from NDP architectures. Likewise, many prior works [1], [36], [51], [54], [55], [145]–[150] observe that applications with high last-level cache MPKI³ are good candidates for NDP.

Figure 1 shows the roofline model (left) and a plot of MPKI vs. speedup (right) of a system with general-purpose NDP support over a baseline system without NDP for a diverse set of 44 applications (see Table 8). In the MPKI vs. speedup plot, the MPKI corresponds to a baseline host CPU system. The speedup represents the performance improvement of a general-purpose NDP system over the baseline (see Section II-D for our methodology). We make the following observations. First, analyzing the roofline model (Figure 1, left), we observe that most of the memory-bound applications (yellow dots) benefit from NDP, as foreseen by prior works. We later observe (in Section III-C1) that such applications are DRAM bandwidth-bound and are a natural fit for NDP. However, the roofline model does *not* accurately account for the NDP suitability of memory-bound applications that (i) benefit from NDP only under particular microarchitectural configurations, e.g., either at low or high core counts (green dots, which are applications that are either bottlenecked by DRAM latency or suffer from L3 cache contention; see Sections III-C3 and III-C4); or (ii) experience performance degradation when executed using NDP (blue dots, which are applications that suffer from the lack of a deep cache hierarchy in NDP architectures; see Section III-C6). Second, analyzing the MPKI vs. speedup plot (Figure 1, right), we observe that while all applications with high MPKI benefit from NDP (yellow dots with MPKI higher than 10), some applications with *low* MPKI can *also* benefit from NDP in all of the NDP microarchitecture configurations we evaluate (yellow dots with MPKI lower than 10) or under specific NDP microarchitecture configurations (green dots with MPKI lower than 10). Thus, even though both the roofline model and MPKI can identify some specific sources

¹We use the term NDP to refer to *any* type of Processing-in-Memory [37].

²See Section II for our experimental evaluation methodology.

³Typically, an MPKI value greater than 10 is considered *high* by prior works [151]–[157].

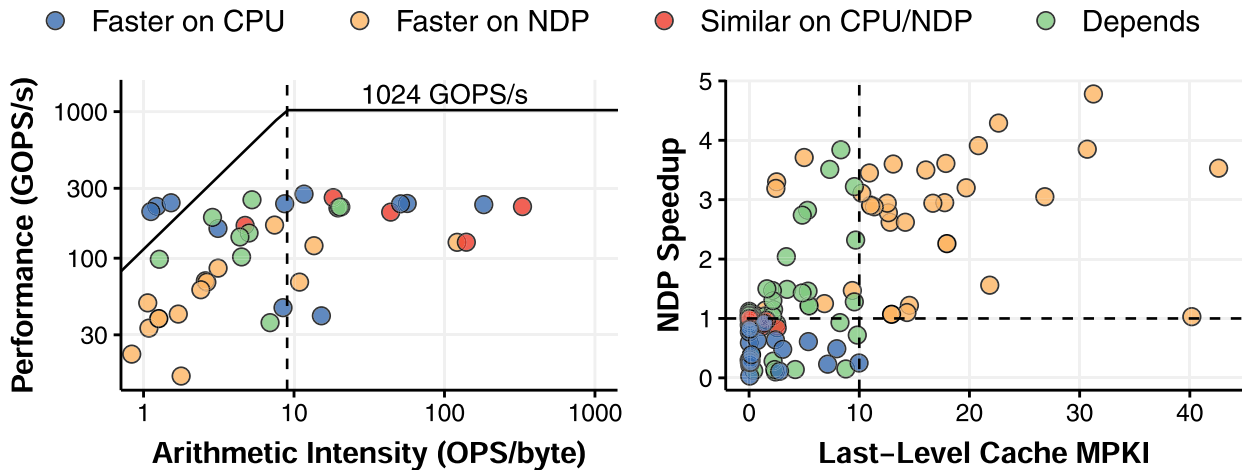


FIGURE 1. Roofline (left) and last-level cache MPKI vs. NDP speedup (right) for 44 memory-bound applications. Applications are classified into four categories: (1) those that experience performance degradation due to NDP (blue; Faster on CPU), (2) those that experience performance improvement due to NDP (yellow; Faster on NDP), (3) those where the host CPU and NDP performance are similar (red; Similar on CPU/NDP), (4) those that experience either performance degradation or performance improvement due to NDP depending on the microarchitectural configuration (green; Depends).

of memory bottlenecks and can sometimes be used as a proxy for NDP suitability, they alone cannot definitively determine NDP suitability because they cannot comprehensively identify different possible sources of memory bottlenecks in a system.

Our goal in this work is (1) to understand the major sources of inefficiency that lead to data movement bottlenecks by observing and identifying relevant metrics and (2) to develop a benchmark suite for data movement that captures each of these sources. To this end, we develop a new three-step methodology to correlate application characteristics with the primary sources of data movement bottlenecks and to determine the potential benefits of three example data movement mitigation mechanisms: (1) a deep cache hierarchy, (2) a hardware prefetcher, and (3) a general-purpose NDP architecture.⁴ We use two main profiling strategies to gather key metrics from applications: (i) an architecture-independent profiling tool and (ii) an architecture-dependent profiling tool. The architecture-independent profiling tool provides metrics that characterize the application memory behavior independently of the underlying hardware. In contrast, the architecture-dependent profiling tool evaluates the impact of the system configuration (e.g., cache hierarchy) on the memory behavior. Our methodology has three steps. In *Step 1*, we use a hardware profiling tool to identify memory-bound functions across many applications. This step allows for a quick first-level identification of many applications that suffer from memory bottlenecks and functions that cause these bottlenecks. In *Step 2*, we use the architecture-independent profiling tool to collect metrics that provide insights about the memory access behavior of the memory-bottlenecked functions. In *Step 3*,

⁴We focus on these three data movement mitigation mechanisms for two different reasons: (1) deep cache hierarchies and hardware prefetchers are standard mechanisms in almost all modern systems, and (2) NDP represents a promising paradigm shift for many modern data-intensive applications.

we collect architecture-dependent metrics and analyze the performance and energy of each function in an application when each of our three candidate data movement mitigation mechanisms is applied to the system. By combining the data obtained from all three steps, we can systematically classify the leading causes of data movement bottlenecks in an application or function into different bottleneck classes.

Using this new methodology, we characterize a large, heterogeneous set of applications (345 applications from 37 different workload suites) across a wide range of domains. Within these applications, we analyze 77K functions and find a subset of 144 functions from 74 different applications that are memory-bound (and that consume a significant fraction of the overall execution time). We fully characterize this set of 144 representative functions to serve as a core set of application kernel benchmarks, which we release as the open-source DAMOV (DAta MOVement) Benchmark Suite [158]. Our analyses reveal six new insights about the sources of memory bottlenecks and their relation to NDP:

- 1) Applications with high last-level cache MPKI and low temporal locality are *DRAM bandwidth-bound*. These applications benefit from the large memory bandwidth available to the NDP system (Section III-C1).
- 2) Applications with low last-level cache MPKI and low temporal locality are *DRAM latency-bound*. These applications do *not* benefit from L2/L3 caches. The NDP system improves performance and energy efficiency by sending L1 misses directly to DRAM (Section III-C2).
- 3) A second group of applications with low LLC MPKI and low temporal locality are *bottlenecked by L1/L2 cache capacity*. These applications benefit from the NDP system at low core counts. However, at high core counts (and thus larger L1/L2 cache space), the caches capture most of the data locality in these

applications, decreasing the benefits the NDP system provides (Section III-C3). We make this observation using a *new* metric that we develop, called *last-to-first miss-ratio (LFMR)*, which we define as the ratio between the number of LLC misses and the total number of L1 cache misses. We find that this metric accurately identifies how efficient the cache hierarchy is in reducing data movement.

- 4) Applications with high temporal locality and low LLC MPKI are *bottlenecked by L3 cache contention* at high core counts. In such cases, the NDP system provides a cost-effective way to alleviate cache contention over increasing the L3 cache capacity (Section III-C4).
- 5) Applications with high temporal locality, low LLC MPKI, and low AI are bottlenecked by the *L1 cache capacity*. The three candidate data movement mitigation mechanisms achieve similar performance and energy consumption for these applications (Section III-C5).
- 6) Applications with high temporal locality, low LLC MPKI, and high AI are *compute-bound*. These applications benefit from a deep cache hierarchy and hardware prefetchers, but the NDP system degrades their performance (Section III-C6).

We publicly release our 144 representative data movement bottlenecked functions from 74 applications as the first open-source benchmark suite for data movement, called DAMOV Benchmark Suite, along with the complete source code for our new characterization methodology [158].

This work makes the following key contributions:

- We propose the first methodology to characterize data-intensive workloads based on the source of their data movement bottlenecks. This methodology is driven by insights obtained from a large-scale experimental characterization of 345 applications from 37 different benchmark suites and an evaluation of the performance of memory-bound functions from these applications with three data-movement mitigation mechanisms.
- We release DAMOV, the first open-source benchmark suite for main memory data movement-related studies, based on our systematic characterization methodology. This suite consists of 144 functions representing different sources of data movement bottlenecks and can be used as a baseline benchmark set for future data-movement mitigation research.
- We show how our DAMOV benchmark suite can aid the study of open research problems for NDP architectures, via four case studies. In particular, we evaluate (i) the impact of load balance and inter-vault communication in NDP systems, (ii) the impact of NDP accelerators on our memory bottleneck analysis, (iii) the impact of different core models on NDP architectures, and (iv) the potential benefits of identifying simple NDP instructions. We conclude that our benchmark suite and methodology can be employed to address many different open research and

development questions on data movement mitigation mechanisms, particularly topics related to NDP systems and architectures.

II. METHODOLOGY OVERVIEW

We develop a new workload characterization methodology to analyze data movement bottlenecks and the suitability of different data movement mitigation mechanisms for these bottlenecks, with a focus on Near-Data Processing (NDP). Our methodology consists of three main steps, as Figure 2 depicts: (1) *memory-bound function identification* using application profiling; (2) *locality-based clustering* to analyze spatial and temporal locality in an architecture-independent manner; and (3) *memory bottleneck classification* using a scalability analysis to nail down the sources of memory boundedness, including architecture-dependent characterization. Our methodology takes as input an application's source code and its input datasets, and produces as output a classification of the primary source of memory bottleneck of important functions in an application (i.e., bottleneck class of each key application function). We illustrate the applicability of this methodology with a detailed characterization of 144 functions that we select from among 77K analyzed functions of 345 characterized applications. In this section, we give an overview of our workload characterization methodology. We use this methodology to drive the analyses we perform in Section III.

A. EXPERIMENTAL EVALUATION FRAMEWORK

As our scalability analysis depends on the hardware architecture, we need a hardware platform that can allow us to replicate and control all of our configuration parameters. Unfortunately, such an analysis cannot be performed practically using real hardware, as (1) there are very few available NDP hardware platforms, and the ones that currently exist do not allow us to comprehensively analyze our general-purpose NDP configuration in a controllable way (as existing platforms are specialized and non-configurable); and (2) the configurations of real CPUs can vary significantly across the range of core counts that we want to analyze, eliminating the possibility of a carefully controlled study. As a result, we must rely on accurate simulation platforms to perform an accurate comparison across different configurations. To this end, we build a framework that integrates the ZSim CPU simulator [159] with the Ramulator memory simulator [160] to produce a fast, scalable, and cycle-accurate open-source simulator called DAMOV-SIM [158]. We use ZSim to simulate the core microarchitecture, cache hierarchy, coherence protocol, and prefetchers. We use Ramulator to simulate the DRAM architecture, memory controllers, and memory accesses. To compute spatial and temporal locality, we modify ZSim to generate a single-thread memory trace for each application, which we use as input for the locality analysis algorithm described in Section II-C (which statically computes the temporal and spatial locality at word-level granularity).

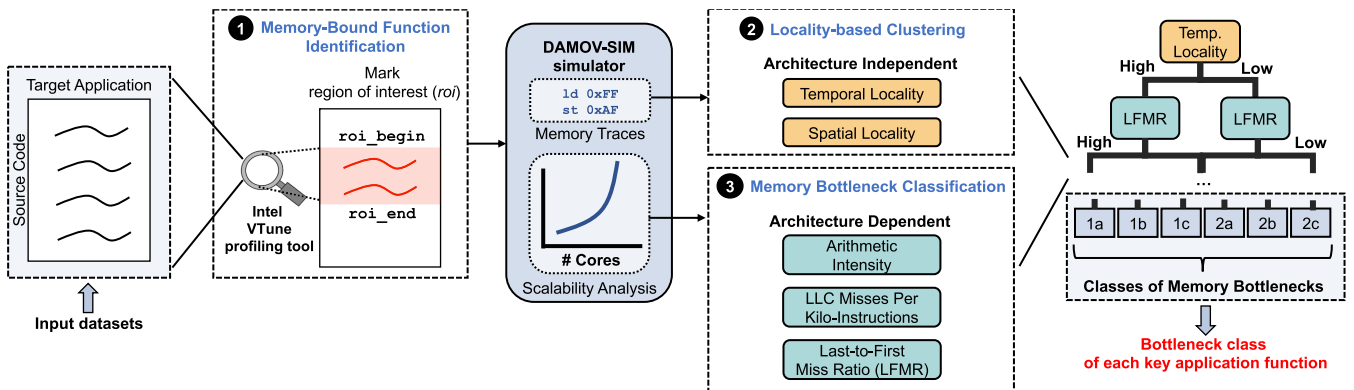


FIGURE 2. Overview of our three-step workload characterization methodology.

B. STEP 1: MEMORY-BOUND FUNCTION IDENTIFICATION

The first step (labeled ① in Figure 2) aims to identify the functions of an application that are *memory-bound* (i.e., functions that suffer from data movement bottlenecks). These bottlenecks might be caused at any level of the memory hierarchy. There are various potential sources of memory boundedness, such as cache misses, cache coherence traffic, and long queuing latencies. Therefore, we need to take all such potential causes into account. This step is optional if the application’s memory-bound functions (i.e., regions of interest, *roi*, in Figure 2) are already known *a priori*.

Hardware profiling tools, both open-source and proprietary, are available to obtain hardware counters and metrics that characterize the application behavior on a computing system. In this work, we use the Intel VTune Profiler [161], which implements the well-known *top-down analysis* [162]. Top-down analysis uses the available CPU hardware counters to hierarchically identify different sources of CPU system bottlenecks for an application. Among the various metrics measured by top-down analysis, there is a relevant one called *Memory Bound* [163] that measures the percentage of CPU pipeline slots that are *not* utilized due to any issue related to data access. We employ this metric to identify functions that suffer from data movement bottlenecks (which we define as functions where *Memory Bound* is greater than 30%).

C. STEP 2: LOCALITY-BASED CLUSTERING

Two key properties of an application’s memory access pattern are its inherent spatial locality (i.e., the likelihood of accessing nearby memory locations in the near future) and temporal locality (i.e., the likelihood of accessing a memory location again in the near future). These properties are closely related to how well the application can exploit the memory hierarchy in computing systems and how accurate hardware prefetchers can be. Therefore, to understand the sources of memory bottlenecks for an application, we should analyze how much spatial and temporal locality its memory accesses inherently exhibit. However, we should isolate these properties from particular configurations of the memory subsystem. Otherwise, it would be unclear if memory bottlenecks are due

to the nature of the memory accesses or due to the characteristics and limitations of the memory subsystem (e.g., limited cache size, too simple or inaccurate prefetching policies). As a result, in this step (labeled ② in Figure 2), we use *architecture-independent* static analysis to obtain spatial and temporal locality metrics for the functions selected in the previous step (Section II-B). Past works [164]–[173] propose different ways of analyzing spatial and temporal locality in an architecture-independent manner. In this work, we use the definition of spatial and temporal metrics presented in [166], [167].

The spatial locality metric is calculated for a window of memory references⁵ of length W using Equation 1. First, for every W memory references, we calculate the minimum distance between any two addresses (*stride*). Second, we create a histogram called the *stride profile*, where each bin i stores how many times each *stride* appears. Third, to calculate the spatial locality, we divide the *percentage* of times *stride* i is referenced (*stride profile*(i)) by the stride length i and sum the resulting value across all instances of i .

$$\text{Spatial Locality} = \sum_{i=1}^{\text{\#bins}} \frac{\text{stride profile}(i)}{i} \quad (1)$$

A spatial locality value close to 0 is caused by large *stride* values (e.g., regular accesses with large strides) or random accesses, while a value equal to 1 is caused by a completely sequential access pattern.

The temporal locality metric is calculated by using a histogram of reused addresses. First, we count the number of times each memory address is repeated in a window of L memory references. Second, we create a histogram called *reuse profile*, where each bin i represents the number of times a memory address is reused, expressed as a power of 2. For each memory address, we increment the bin that represents the corresponding number of repetitions. For example, *reuse profile*(0) represents memory addresses that are reused only once. *reuse profile*(1) represents memory addresses that are

⁵We compute both the spatial and temporal locality metrics at the word granularity. In this way, we keep our locality analysis architecture-independent, using *only* properties of the application under study.

reused twice. Thus, if a memory address is reused N times, we increment $reuse\ profile(\lfloor \log_2 N \rfloor)$ by one. Third, we obtain the temporal locality metric with Equation 2.

$$Temporal\ Locality = \sum_{i=0}^{\#bins} \frac{2^i \times reuse\ profile(i)}{total\ memory\ accesses} \quad (2)$$

A temporal locality value of 0 indicates no data reuse, while a value close to 1 indicates very high data reuse (i.e., a value equal to 1 means that the application accesses a single memory address continuously).

To calculate these metrics, we empirically select window lengths W and L to 32. We find that different values chosen for W and L do not significantly change the conclusions of our analysis. We observe that our conclusions remain the same when we set both values to 8, 16, 32, 64, and 128.

D. STEP 3: BOTTLENECK CLASSIFICATION

While Step 2 allows us to understand inherent application sources for memory boundedness, it is important to understand how hardware architectural features can also result in memory bottlenecks. As a result, in our third step (3 in Figure 2), we perform a scalability analysis of the functions selected in Step 1, where we evaluate performance and energy scaling for three different system configurations. The scalability analysis makes use of three *architecture-dependent* metrics: (1) *Arithmetic Intensity (AI)*, (2) *Misses per Kilo-Instruction (MPKI)*, and (3) a new metric called *Last-to-First Miss-Ratio (LFMR)*. We select these metrics for the following reasons. First, AI can measure the compute intensity of an application. Intuitively, we expect an application with high compute intensity to not suffer from severe data movement bottlenecks, as demonstrated by prior work [174]. Second, MPKI serves as a proxy for the memory intensity of an application. It can also indicate the memory pressure experienced by the main memory system [45], [47], [48], [58], [151], [153], [156], [175]–[177]. Third, LFMR, a new metric we introduce and is described in detail later in this subsection, indicates how efficient the cache hierarchy is in reducing data movement.

As part of our methodology development, we evaluate other metrics related to data movement, including raw cache misses, coherence traffic, and DRAM row misses/hits/conflicts. We observe that even though such metrics are useful for further characterizing an application (as we do in some of our later analyses in Section III-C), they do not necessarily characterize a specific type of data movement bottleneck. We show in Section IV-A that the three architecture-dependent and two architecture-independent metrics we select for our classification are enough to accurately characterize and cluster the different types of data movement bottlenecks in a wide variety of applications.

1) DEFINITION OF METRICS

We define Arithmetic Intensity (AI) as the number of arithmetic and logic operations performed per L1 cache line

accessed.⁶ This metric indicates how much computation there is per memory request. Intuitively, applications with high AI are likely to be computationally intensive, while applications with low AI tend to be memory intensive. We use MPKI at the last-level cache (LLC), i.e., the number of LLC misses per one thousand instructions. This metric is considered to be a good indicator of NDP suitability by several prior works [1], [36], [51], [54], [55], [145]–[149]. We define the LFMR of an application as the ratio between the number of LLC misses and the total number of L1 cache misses. We find that this metric accurately identifies how much an application benefits from the deep cache hierarchy of a contemporary CPU. An LFMR value close to 0 means that the number of LLC misses is very small compared to the number of L1 misses, i.e., the L1 misses are likely to hit in the L2 or L3 caches. However, an LFMR value close to 1 means that very few L1 misses hit in L2 or L3 caches, i.e., the application does not benefit much from the deep cache hierarchy, and most L1 misses need to be serviced by main memory.

2) SCALABILITY ANALYSIS AND SYSTEM CONFIGURATION

The goal of the scalability analysis we perform is to nail down the specific sources of data movement bottlenecks in the application. In this analysis, we (i) evaluate the performance and energy scaling of an application in three different system configurations; and (ii) collect the key metrics for our bottleneck classification (i.e., AI, MPKI, and LFMR). During scalability analysis, we simulate three system configurations of a general-purpose multicore processor:

- A host CPU with a deep cache hierarchy (i.e., private L1 (32 kB) and L2 (256 kB) caches, and a shared L3 (8 MB) cache with 16 banks). We call this configuration *Host CPU*.
- A host CPU with a deep cache hierarchy (same cache configurations as in *Host CPU*), augmented with a stream prefetcher [178]. We call this configuration *Host CPU with prefetcher*.
- An NDP CPU with a single level of cache (only a private read-only⁷ L1 cache (32 kB), as assumed in many prior NDP works [1], [46], [51], [63], [66], [74], [99], [101], [119], [179]) and no hardware prefetcher. We call this configuration *NDP*.

The remaining components of the processor configuration are kept the same (e.g., number of cores, instruction window size, branch predictor) to isolate the impact of only the caches, prefetchers, and NDP. This way, we expect that the performance and energy differences between the

⁶We consider AI to be architecture-dependent since we consider the number of cache lines accessed by the application (and hence the hardware cache block size) to compute the metric. This is the same definition of AI used by the hardware profiling tool we employ in Step 1 (i.e., the Intel VTune Profiler [161]).

⁷We use read-only L1 caches to simplify the cache coherence model of the NDP system. Enabling efficient synchronization and cache coherence in NDP architectures is an open-research problem, as we discuss in Section III-F.

three configurations to come *exclusively* from the different data movement requirements. For the three configurations, we sweep the number of CPU cores in our analysis from 1 to 256, as previous works [46], [66], [180] show that large core counts are necessary to saturate the bandwidth provided by modern high-bandwidth memories, and because modern CPUs and NDP proposals can have varying core counts. The core count sweep allows us to observe (1) how an application's performance changes when increasing the pressure on the memory subsystem, (2) how much Memory-Level Parallelism (MLP) [176], [181]–[184] the application has, and (3) how much the cores leverage the cache hierarchy and the available memory bandwidth. We proportionally increase the size of the CPU's private L1 and L2 caches when increasing the number of CPU cores in our analysis (e.g., when scaling the CPU core count from 1 to 4, we also scale the aggregated L1/L2 cache size by a factor of 4). We use out-of-order and in-order CPU cores in our analysis for all three configurations. In this way, we build confidence that our trends and findings are independent of a specific underlying general-purpose core microarchitecture. We simulate a memory architecture similar to the Hybrid Memory Cube (HMC) [73], where (1) the host CPU accesses memory through a high-speed off-chip link, and (2) the NDP logic resides in the logic layer of the memory chip and has direct access to the DRAM banks (thus taking advantage of higher memory bandwidth and lower memory latency). Table 1 lists the parameters of our host CPU, host CPU with prefetcher, and NDP baseline configurations.

3) CHOOSING AN NDP ARCHITECTURE

We note that across the proposed NDP architectures in literature, there is a lack of consensus on whether the architectures should make use of general-purpose NDP cores or specialized NDP accelerators [36], [37]. In this work, we focus on general-purpose NDP cores for two major reasons. First, many prior works (e.g., [1], [46], [51], [63], [66], [76], [99], [101], [119], [147], [179], [190], [192]–[194]) suggest that general-purpose cores (especially simple in-order cores) can successfully accelerate memory-bound applications in NDP architectures. In fact, UPMEM [83], a start-up building some of the first commercial in-DRAM NDP systems, utilizes simple in-order cores in their NDP units inside DRAM chips [83], [140]. Therefore, we believe that general-purpose NDP cores are a promising candidate for future NDP architectures. Second, the goal of our work is not to perform a design space exploration of different NDP architectures, but rather to understand the key properties of applications that lead to memory bottlenecks that can be mitigated by a simple NDP engine. While we expect that each application could potentially benefit further from an NDP accelerator tailored to its computational and memory requirements, such customized architectures open many challenges for a methodical characterization, such as the need for significant code refactoring, changes in data mapping,

and code partitioning between NDP accelerators and host CPUs.^{8,9}

III. CHARACTERIZING MEMORY BOTTLENECKS

In this section, we apply our three-step workload characterization methodology to characterize the sources of memory bottlenecks across a wide range of applications. First, we apply *Step 1* to identify memory-bound functions within an application (Section III-A). Second, we apply *Step 2* and cluster the identified functions using two architecture-independent metrics (spatial and temporal locality) (Section III-B). Third, we apply *Step 3* and combine the architecture-dependent and architecture-independent metrics to classify the different sources of memory bottlenecks we observe (Section III-C).

We also evaluate various other aspects of our three-step workload characterization methodology. We investigate the effect of increasing the last-level cache on our memory bottleneck classification in Section III-D. We provide a validation of our memory bottleneck classification in Section III-E. We discuss the limitations of our proposed methodology in Section III-F.

A. STEP 1: MEMORY-BOUND FUNCTION IDENTIFICATION

We first apply Step 1 of our methodology across 345 applications (listed in Appendix C) to identify functions whose performance is significantly affected by data movement. We use the previously-proposed top-down analysis methodology [162] that has been used by several recent workload characterization studies [5], [195], [196]. As discussed in Section II-B, we use the Intel VTune Profiler [161], which we run on an Intel Xeon E3-1240 processor [197] with four cores. We disable hyper-threading for more accurate profiling results, as recommended by the VTune documentation [198]. For the applications that we analyze, we select functions (1) that take at least 3% of the clock cycles, and (2) that have a Memory Bound percentage that is greater than 30%. We choose 30% as the threshold for this metric because, in preliminary simulation experiments, we do not observe significant performance improvement or energy savings with data movement mitigation mechanisms for functions whose Memory Bound percentage is less than 30%.

The applications we analyze come from a variety of sources, such as popular workload suites (Chai [199], CORAL [200], Parboil [201], PARSEC [202], Rodinia [203], SD-VBS [204], SPLASH-2 [205]), benchmarking (STREAM [120], HPCG [206], HPCG [207]), bioinformatics [208], databases [209], [210], graph processing frameworks (GraphMat [211], Ligra [212]), a map-reduce framework (Phoenix [213]), and neural

⁸We show in Section V-B that our DAMOV benchmark suite is useful to rigorously study NDP accelerators.

⁹The development of a new methodology or extension of our methodology to perform analysis targeting function-specific, customized, or reconfigurable NDP accelerators is a good direction for future work.

TABLE 1. Evaluated Host CPU and NDP system configurations.

Host CPU Configuration	
Host CPU Processor	1, 4, 16, 64, and 256 cores @2.4 GHz, 32 nm; 4-wide out-of-order 1, 4, 16, 64, and 256 cores @2.4 GHz, 32 nm; 4-wide in-order Buffers: 128-entry ROB; 32-entry LSQ (each) Branch predictor: Two-level GAs [185], 2,048 entry BTB; 1 branch per fetch
Private L1 Cache	32 KB, 8-way, 4-cycle; 64 B line; LRU policy Energy: 15/33 pJ per hit/miss [51, 186]
Private L2 Cache	256 KB, 8-way, 7-cycle; 64 B line; LRU policy MSHR size: 20-request, 20-write, 10-eviction Energy: 46/93 pJ per hit/miss [51, 186]
Shared L3 Cache	8 MB (16-banks), 0.5 MB per bank, 16-way, 27-cycle 64 B line; LRU policy; Bi-directional ring [187]; Inclusive; MESI protocol [188] MSHR size: 64-request, 64-write, 64-eviction Energy: 945/1904 pJ per hit/miss [51, 186]
Host CPU with Prefetcher Configuration	
Processor, Private L1 Cache, Private L2 Cache, and Share L3 Cache	Same as in Host CPU Configuration
L2 Cache Prefetcher	Stream prefetcher [178, 189]; 2-degree; 16 stream buffers; 64 entries
NDP Configuration	
NDP CPU Processor	1, 4, 16, 64, and 256 cores @2.4 GHz, 32 nm; 4-wide out-of-order 1, 4, 16, 64, and 256 cores @ 2.4 GHz, 32 nm; 4-wide in-order Buffers: 128-entry ROB; 32-entry LSQ (each) Branch predictor: Two-level GAs [185], 2,048 entry BTB; 1 branch per fetch
Private L1 Cache	32 KB, 8-way, 4-cycle; 64 B line; LRU policy; Read-only Data Cache Energy: 15/33 pJ per hit/miss [51, 186]
Common	
Main Memory	HMC v2.0 Module [73] 32 vaults, 8 DRAM banks/vault, 256 B row buffer 8 GB total size; DRAM@166 MHz; 4-links@8 GHz 8 B burst width at 2:1 core-to-bus freq. ratio; Open-page policy; HMC default interleaving [45, 73] ¹⁰ Energy: 2 pJ/bit internal, 8 pJ/bit logic layer [51, 64, 190], 2 pJ/bit links [51, 76, 191]

networks (AlexNet [214], Darknet [215]). We explore different input dataset sizes for the applications and choose real-world input datasets that impose high pressure on the memory subsystem (as we expect that such real-world inputs are best suited for stressing the memory hierarchy). We also use different inputs for applications whose performance is tightly related to the input dataset properties. For example, we use two different graphs with varying connectivity degrees (rMat [217] and USA [218]) to evaluate graph processing applications and two different read sequences to evaluate read alignment algorithms [60], [219], [220].

In total, our application analysis covers more than 77K functions. To date, this is the most extensive analysis of data movement bottlenecks in real-world applications. We find a set of 144 functions that take at least 3% of the total clock cycles and have a value of the Memory Bound metric greater or equal to 30%, which forms the basis of DAMOV, our

¹⁰The default HMC interleaving scheme (Row:Column:Bank:Vault [73]) interleaves consecutive cache lines across vaults, and then across banks [216].

new data movement benchmark suite. We provide a list of all 144 functions selected based on our analysis and their major characteristics in Appendix A.

After identifying memory-bound functions over a wide range of applications, we apply Steps 2 and 3 of our methodology to classify the primary sources of memory bottlenecks for our selected functions. We evaluate a total of 144 functions out of the 77K functions we analyze in Step 1. These functions span across 74 different applications, belonging to 16 different widely-used benchmark suites or frameworks.

From the 144 functions that we analyze further, we select a subset of 44 representative functions to explore in-depth in Sections III-B and III-C and to drive our bottleneck classification analysis. We use the 44 representative functions to ease our explanations and make figures more easily readable. Table 8 in Appendix A lists the 44 representative functions that we select. The table includes one column that indicates the class of data movement bottleneck experienced by each function (we discuss the classes in Section III-C), and another column representing the percentage of clock cycles of the selected function in the whole application. We select

representative functions that belong to a variety of domains: benchmarking, bioinformatics, data analytics, databases, data mining, data reorganization, graph processing, neural networks, physics, and signal processing. In Section III-E, we validate our classification using the remaining 100 functions and provide a summary of the results of our methodology when applied to all 144 functions.

B. STEP 2: LOCALITY-BASED CLUSTERING

We cluster the 44 representative functions across both spatial and temporal locality using the K-means clustering algorithm [221]. Figure 3 shows how each function is grouped. We find that two groups emerge from the clustering: (1) low temporal locality functions (orange boxes in Figure 3), and (2) high temporal locality functions (blue boxes in Figure 3). Intuitively, the closer a function is to the bottom-left corner of the figure, the less likely it is to take advantage of a multi-level cache hierarchy. These functions are more likely to be good candidates for NDP. However, as we see in Section III-C, the NDP suitability of a function also depends on a number of other factors.

C. STEP 3: BOTTLENECK CLASSIFICATION

Within the two groups of functions identified in Section III-B, we use three key metrics (AI, MPKI, and LFMR) to classify the memory bottlenecks. We observe that the AI of the analyzed low temporal locality functions is low (i.e., always less than 2.2 ops/cache line, with an average of 1.3 ops/cache

line). Among the high temporal locality functions, there are some with low AI (minimum of 0.3 ops/cache line) and others with high AI (maximum of 44 ops/cache line). LFMR indicates whether a function benefits from a deeper cache hierarchy. When LFMR is low (i.e., less than 0.1), then a function benefits significantly from a deeper cache hierarchy, as most misses from the L1 cache hit in either the L2 or L3 caches. When LFMR is high (i.e., greater than 0.7), then most L1 misses are not serviced by the L2 or L3 caches, and must go to memory. A medium LFMR (0.1–0.7) indicates that a deeper cache hierarchy can mitigate some, but not a very large fraction of L1 cache misses. MPKI indicates the memory intensity of a function (i.e., the rate at which requests are issued to DRAM). We say that a function is memory-intensive (i.e., it has a high MPKI) when the MPKI is greater than 10, which is the same threshold used by prior works [151]–[157].

We find that six classes of functions emerge, based on their temporal locality, AI, MPKI, and LFMR values, as we observe from Figures 3 and 4. We observe that spatial locality is not a key metric for our classification (i.e., it does not define a bottleneck class) because the L1 cache, which is present in both host CPU and NDP system configurations, can capture most of the spatial locality for a function. Figure 4 shows the LFMR and MPKI values for each class. Note that we do not have classes of functions for all possible combinations of metrics. In our analysis, we obtain the temporal locality, AI, MPKI, and LFMR values and their combinations empirically.

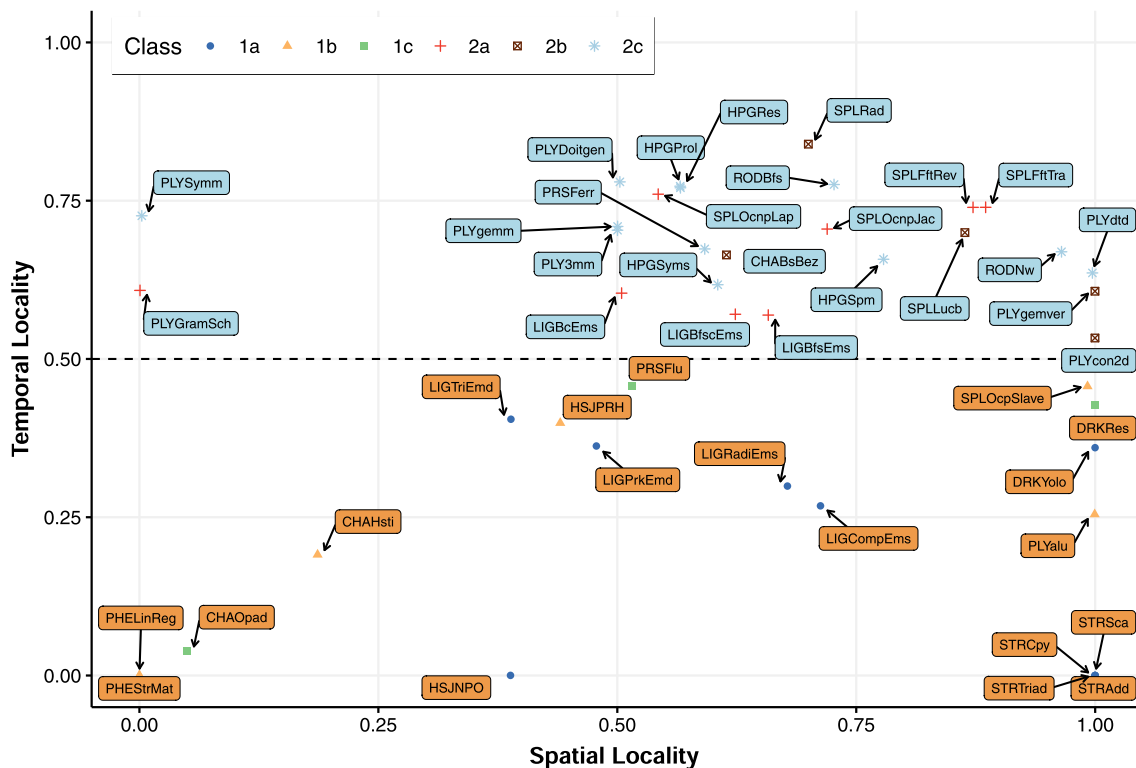


FIGURE 3. Locality-based clustering of 44 representative functions.

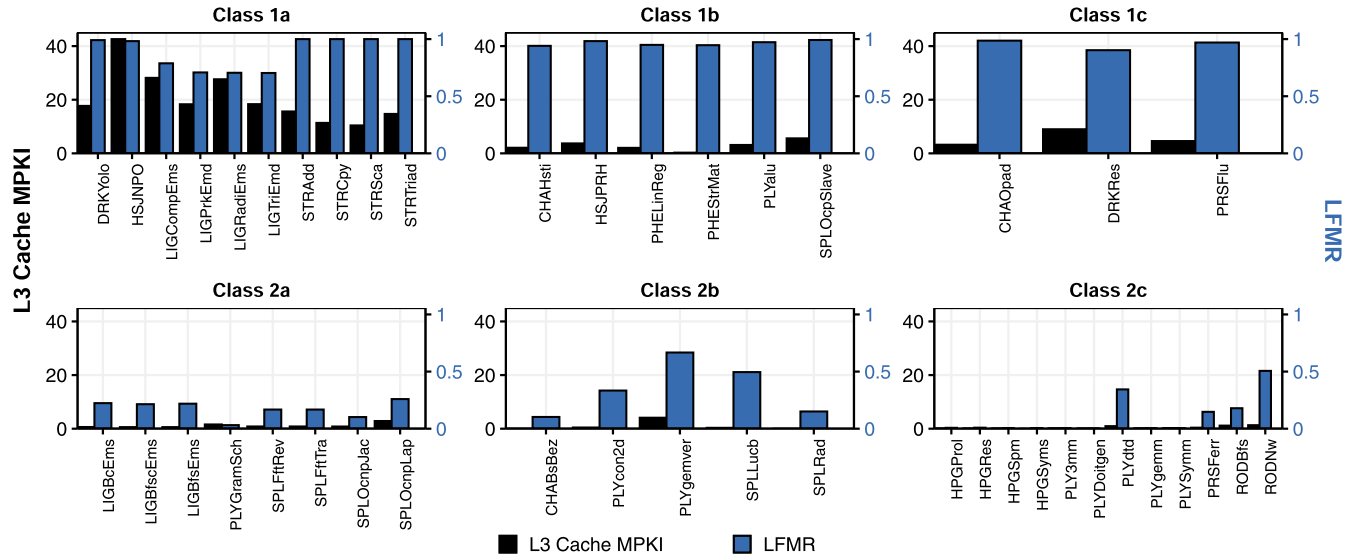


FIGURE 4. L3 Cache MPKI and Last-to-First Miss Ratio (LFMR) for 44 representative functions.

Fundamentally, not all value combinations of different metrics are possible. We list some of the combinations we do *not* observe in our analysis of 144 functions:

- A function with high LLC MPKI does *not* display low LFMR. This is because a low LFMR happens when most L1 misses hit the L2/L3 caches. Thus, it becomes highly unlikely for the L3 cache to suffer many misses when the L2/L3 caches do a good job in fulfilling L1 cache misses.
- A function with high temporal locality does *not* display both high LFMR and high MPKI. This is because a function with high temporal locality will likely issue repeated memory requests to few memory addresses, which will likely be serviced by the cache hierarchy.
- A function with low temporal locality does *not* display low LFMR since there is little data locality to be captured by the cache hierarchy.

We discuss each class in detail below, identifying the memory bottlenecks for each class and whether the NDP system can alleviate these bottlenecks. To simplify our explanations, we focus on a smaller set of 12 representative functions (out of the 44 representative functions) for this part of the analysis. Figure 5 shows how each of the 12 functions scales in terms of performance for the *host CPU*, *host CPU with prefetcher*, and *NDP* system configurations.

1) CLASS 1a: LOW TEMPORAL LOCALITY, LOW AI, HIGH LFMR, AND HIGH MPKI (DRAM BANDWIDTH-BOUND FUNCTIONS)

Functions in this class exert high main memory pressure since they are highly memory intensive and have low data reuse. To understand how this affects a function's suitability for NDP, we study how performance scales as we increase the number of cores available to a function, for the host CPU, host CPU with prefetcher, and NDP system configurations.

Figure 5(a) depicts performance¹¹ as we increase the core count, normalized to the performance of one host CPU core, for two representative functions from Class 1a (HSJNPO and LIGPrkEms); we see similar trends for all functions in the class.

We make three observations from the figure. First, as the number of host CPU cores increases, performance eventually stops increasing significantly. For HSJNPO, host CPU performance increases by $27.5\times$ going from 1 to 64 host CPU cores but only 27% going from 64 host CPU cores to 256 host CPU cores. For LIGPrkEms, host CPU performance increases by $33\times$ going from 1 to 64 host CPU cores but *decreases* by 20% going from 64 to 256 host CPU cores. We find that the lack of performance improvement at large host CPU core counts is due to main memory bandwidth saturation, as shown in Figure 6. Given the limited DRAM bandwidth available across the off-chip memory channel, we find that Class 1a functions saturate the DRAM bandwidth once enough host CPU cores (e.g., 64) are used, and thus these functions are *bottlenecked by the DRAM bandwidth*. Second, the host CPU system with prefetcher slows down the execution of the HSJNPO (LIGPrkEms) function compared with the host CPU system without prefetcher by 43% (38%), on average across all core counts. The prefetcher is ineffective since these functions have low temporal and spatial locality. Third, when running on the NDP system, the functions see continued performance improvements as the number of NDP cores increases. By providing the functions with access to the much higher bandwidth available inside memory, the NDP system can greatly outperform the host CPU system at a high enough core count. For example, at 64/256 cores, the NDP system outperforms the host CPU system by $1.7\times/4.8\times$ for HSJNPO, and by $1.5\times/4.1\times$ for LIGPrkEms.

¹¹Performance is the inverse of application execution time.

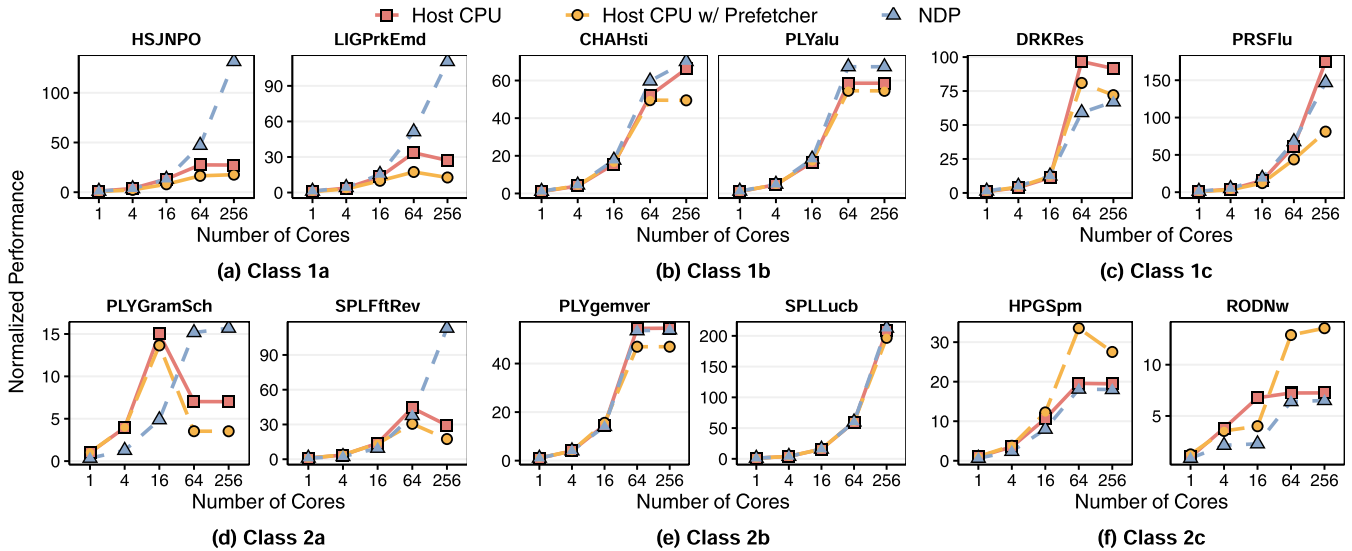


FIGURE 5. Performance of 12 representative functions on three systems: host CPU, host CPU with prefetcher, and NDP, normalized to one host CPU core.

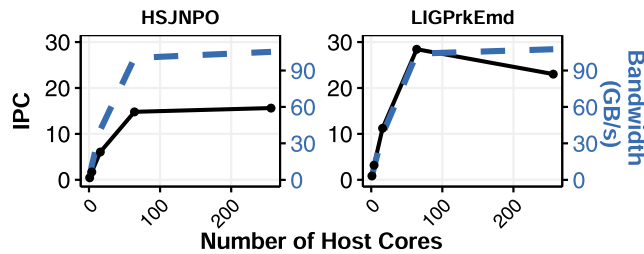


FIGURE 6. Host CPU system IPC vs. utilized DRAM Bandwidth for representative Class 1a functions.

Figure 7 depicts the energy breakdown for our two representative functions. We make two observations from the figure. First, for HSJNPO, the energy spent on DRAM for both host CPU system and NDP system are similar. This is due to the function’s poor locality, as 98% of its memory requests miss in the L1 cache. Since LFMR is near 1, L1 miss requests almost always miss in the L2 and L3 caches and go to DRAM in the host CPU system for all core counts we evaluate, which requires significant energy to query the large caches and then to perform off-chip data transfers. The NDP system does not access L2, L3, and off-chip links, leading to large system energy reduction. Second, for LIGPrkEmd, the DRAM energy is higher in the NDP system than in the host CPU system. Since the function’s LFMR is 0.7, some memory requests that would be cache hits in the host CPU’s L2 and L3 caches are instead sent directly to DRAM in the NDP system. However, the total energy consumption on the host CPU system is still larger than that on the NDP system, again because the NDP system eliminates the L2, L3 and off-chip link energy.

DRAM bandwidth-bound applications such as those in Class 1a have been the primary focus of a large number of proposed NDP architectures (e.g., [1], [46], [54], [69], [76], [132], [133], [192], [222], [223]), as they benefit from increased main memory bandwidth and do

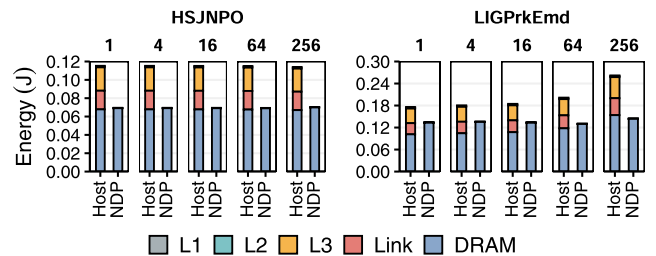


FIGURE 7. Cache and DRAM energy breakdown for representative Class 1a functions at 1, 4, 16, 64, and 256 cores.

not have high AI (and, thus, do not benefit from complex cores on the host CPU system). An NDP architecture for a function in Class 1a needs to extract enough MLP [57], [176], [181]–[184], [224]–[229] to maximize the usage of the available internal memory bandwidth. However, prior work has shown that this can be challenging due to the area and power constraints in the logic layer of a 3D-stacked DRAM [1], [46]. To exploit the high memory bandwidth while satisfying these area and power constraints, the NDP architecture should leverage application memory access patterns to efficiently maximize main memory bandwidth utilization.

We find that there are two dominant types of memory access patterns among our Class 1a functions. First, functions with regular access patterns (DRKYolo, STRAdd, STRCpy, STRSca, STRTriad) can take advantage of specialized accelerators or Single Instruction Multiple Data (SIMD) architectures [1], [66], which can exploit the regular access patterns to issue many memory requests concurrently. Such accelerators or SIMD architectures have hardware area and thermal dissipation that fall well within the constraints of 3D-stacked DRAM [1], [46], [64], [230]. Second, functions with irregular access patterns (HSJNPO, LIGCompEms, LIGPrkEmd, LIGRadiEms) require techniques to extract MLP while still fitting within the design constraints. This

requires techniques that cater to the irregular memory access patterns, such as prefetching algorithms designed for graph processing [46], [231]–[235], pre-execution of difficult access patterns [57], [58], [151], [183], [184], [236]–[243] or hardware accelerators for pointer chasing [55], [56], [149], [193], [244]–[246].

2) CLASS 1b: LOW TEMPORAL LOCALITY, LOW AI, HIGH LFMR, AND LOW MPKI (DRAM LATENCY-BOUND FUNCTIONS)

While functions in this class do not effectively use the host CPU caches, they do *not* exert high pressure on the main memory due to their low MPKI. Across all Class 1b functions, the average DRAM bandwidth consumption is only 0.5 GB/s. However, all the functions have very high LFMR values (the minimum is 0.94 for CHAHsti), indicating that the host CPU L2 and L3 caches are ineffective. Because the functions cannot exploit significant MLP but still incur long-latency requests to DRAM, the DRAM requests fall on the critical path of execution and stall forward progress [57], [58], [151], [176], [247]. Thus, Class 1b functions are *bottlenecked by DRAM latency*. Figure 5(b) shows performance of both the host CPU system and the NDP system for two representative functions from Class 1b (CHAHsti and PLYalu). We observe that while performance of both the host CPU system and the NDP system scale well as the core count increases, NDP system performance is always higher than the host CPU system performance for the same core count. The maximum (average) speedup with NDP over host CPU at the same core count is $1.15\times$ ($1.12\times$) for CHAHsti and $1.23\times$ ($1.13\times$) for PLYalu.

We find that the NDP system’s improved performance is due to a reduction in the Average Memory Access Time (AMAT) [248]. Figure 8 shows the AMAT for our two representative functions. Memory accesses take significantly longer in the host CPU system than in the NDP system due to the additional latency of looking up requests in the L2 and L3 caches, even though data is rarely present in those caches, and going through the off-chip links.

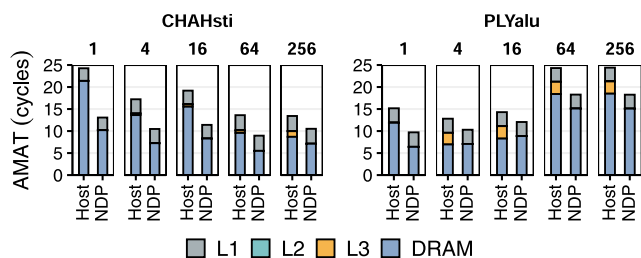


FIGURE 8. Average Memory Access Time (AMAT) for representative Class 1b functions.

Figure 9 shows the energy breakdown for Class 1b representative functions. Similar to Class 1a, we observe that the L2/L3 caches and off-chip links are a large source of energy usage in the host CPU system. While DRAM energy increases in the NDP system, as L2/L3 hits in the host CPU system become DRAM lookups with NDP, the overall energy

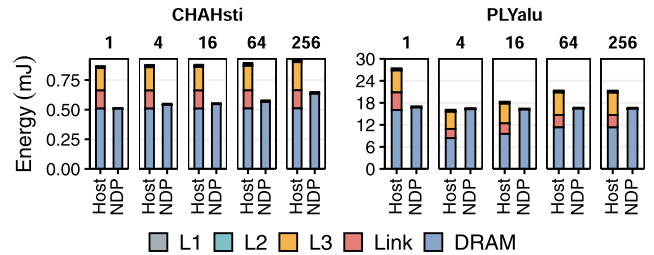


FIGURE 9. Energy breakdown for representative Class 1b functions.

consumption in the NDP system is greatly smaller (by 69% maximum and 39% on average) due to the lack of L2 and L3 caches.

Class 1b functions benefit from the NDP system, but primarily because of the lower memory access latency (and energy) that the NDP system provides for memory requests that need to be serviced by DRAM. These functions could benefit from other latency and energy reduction techniques, such as L2/L3 cache bypassing [51], [249]–[260], low-latency DRAM [15], [22]–[26], [89], [127], [261]–[276], and better memory access scheduling [153]–[157], [175]–[177], [247], [277]–[290]. However, they generally do *not* benefit significantly from prefetching (as seen in Figure 5(b)), since infrequent memory requests make it difficult for the prefetcher to successfully train on an access pattern.

3) CLASS 1c: LOW TEMPORAL LOCALITY, LOW AI, DECREASING LFMR WITH CORE COUNT, AND LOW MPKI (L1/L2 CACHE CAPACITY BOTTLENECKED FUNCTIONS)

We find that the behavior of functions in this class depends on the number of cores they are using. Figure 5(c) shows the host CPU system and the NDP system performance as we increase the core count for two representative functions (DRKRes and PRSF1u). We make two observations from the figure. First, at low core counts, the NDP system outperforms the host CPU system. With a low number of cores, the functions have medium to high LFMR (0.5 for DRKRes at 1 and 4 host CPU cores; 0.97 at 1 host CPU core and 0.91 at 4 host CPU cores for PRSF1u), and behave like Class 1b functions, where they are DRAM latency-sensitive. Second, as the core count increases, the host CPU system begins to outperform the NDP system. For example, beyond 16 (64) cores, the host CPU system outperforms the NDP system for DRKRes (PRSF1u). This is because as the core count increases, the aggregate L1 and L2 cache size available at the host CPU system grows, which reduces the miss rates of both L2 and L3 caches. As a result, the LFMR decreases significantly (e.g., at 256 cores, LFMR is 0.09 for DRKRes and 0.35 for PRSF1u). This indicates that the *available L1/L2 cache capacity* bottlenecks Class 1c functions.

Figure 10 shows the energy breakdown for Class 1c functions. We make three observations from the figure. First, for functions with larger LFMR values (PRSF1u), the NDP system provides energy savings over the host CPU system

at lower core counts, since the NDP system eliminates the energy consumed due to L3 and off-chip link accesses. Second, for functions with smaller LFMR values (DRKRes), the NDP system does not provide energy savings even for low core counts. Due to the medium LFMR, enough requests still hit in the host CPU system L2/L3 caches, and these cache hits become DRAM accesses in the NDP system, which consume more energy than the cache hits. Third, at high-enough core counts, the NDP system consumes more energy than the host CPU system for all Class 1c functions. As the LFMR decreases, the functions effectively utilize the caches in the host CPU system, reducing the off-chip traffic and, consequently, the energy Class 1c functions spend on accessing DRAM. The NDP system, which does not have L2 and L3 caches, pays the larger energy cost of a DRAM access for all L2/L3 hits in the host CPU system.

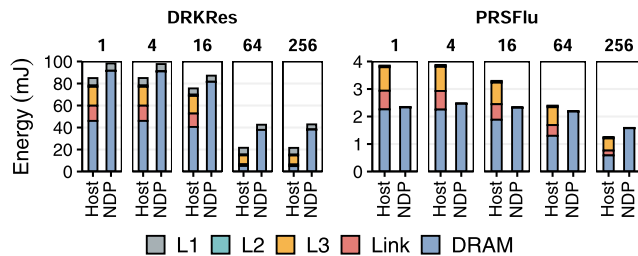


FIGURE 10. Energy breakdown for representative Class 1c functions.

We find that the primary source of the memory bottleneck in Class 1c functions is limited L1/L2 cache capacity. Therefore, while the NDP system improves performance and energy of some Class 1c functions at low core counts (with lower associated L1/L2 cache capacity), the NDP system does not provide performance and energy benefits across all core counts for Class 1c functions.

4) CLASS 2a: HIGH TEMPORAL LOCALITY, LOW AI, INCREASING LFMR WITH CORE COUNT, AND LOW MPKI (L3 CACHE CONTENTION BOTTLENECKED FUNCTIONS)

Like Class 1c functions, the behavior of the functions in this class depends on the number of cores that they use. Figure 5(d) shows the host CPU system and the NDP system performance as we increase the core count for two representative functions (PLYGramSch and SPLFftRev). We make two observations from the figure. First, at low core counts, the functions do *not* benefit from the NDP system. In fact, for a single core (16 cores), PLYGramSch *slows down* by 67% ($3\times$) when running on the NDP system, compared to running on the host CPU system. This is because, at low core counts, these functions make reasonably good use of the cache hierarchy, with LFMR values of 0.03 for PLYGramSch and lower than 0.44 for SPLFftRev until 16 host CPU cores. We confirm this in Figure 11, where we see that very few memory requests for PLYGramSch and SPLFftRev go to DRAM (5% for PLYGramSch, and at most 13% for SPLFftRev) at core counts lower than 16. Second, at high core counts (i.e., 64 for PLYGramSch and

256 for SPLFftRev), the host CPU system performance starts to *decrease*. This is because Class 2a functions are *bottlenecked by cache contention*. At 256 cores, this contention undermines the cache effectiveness and causes the LFMR to increase to 0.97 for PLYGramSch and 0.93 for SPLFftRev. With the last-level cache rendered essentially ineffective, the NDP system greatly improves performance over the host CPU system: by $2.23\times$ for PLYGramSch and $3.85\times$ for SPLFftRev at 256 cores.

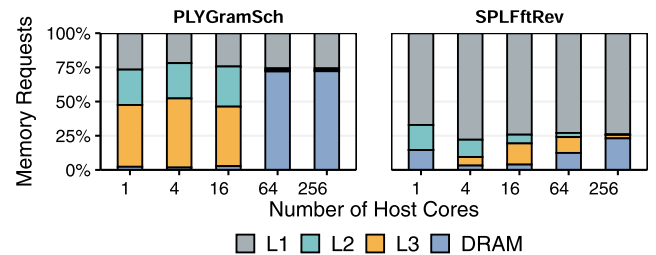


FIGURE 11. Memory request breakdown for representative Class 2a functions.

One impact of the increased cache contention is that it converts these high-temporal-locality functions into memory latency-bound functions. We find that with the increased number of requests going to DRAM due to cache contention, the AMAT increases significantly, in large part due to queuing at the memory controller. At 256 cores, the queuing becomes so severe that a large fraction of requests (24% for PLYGramSch and 67% for SPLFftRev) must be reissued because the memory controller queues are full. The increased main memory bandwidth available to the NDP cores allows the NDP system to issue many more requests concurrently, which reduces the average length of the queue and, thus, the main memory latency. The NDP system also reduces memory access latency by getting rid of L2/L3 cache lookup and interconnect latencies.

Figure 12 shows the energy breakdown for the two representative Class 2a functions. We make two observations. First, the host CPU system is more energy-efficient than the NDP system at low core counts, as most of the memory requests are served by on-chip caches in the host CPU system. Second, the NDP system provides large energy savings over the host CPU system at high core counts. This is due to the increased cache contention, which increases the number of off-chip requests that the host CPU

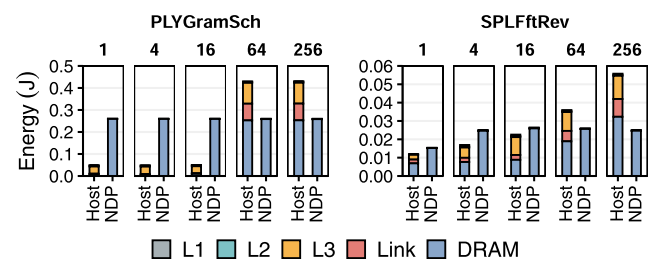


FIGURE 12. Energy breakdown for representative Class 2a functions.

system must make, increasing the L3 and off-chip link energy.

We conclude that cache contention is the primary scalability bottleneck for Class 2a functions, and the NDP system can provide an effective way of mitigating this cache contention bottleneck without incurring the high area and energy overheads of providing additional cache capacity in the host CPU system, thereby improving the scalability of these applications to high core counts.

5) CLASS 2b: HIGH TEMPORAL LOCALITY, LOW AI, LOW/MEDIUM LFMR, AND LOW MPKI (L1 CACHE CAPACITY BOTTLENECKED FUNCTIONS)

Figure 5(e) shows the host CPU system and the NDP system performance for PLYgemver and SPLLucb. We make two observations from the figure. First, as the number of cores increases, performance of the host CPU system and the NDP system scale in a very similar fashion. The NDP system and the host CPU system perform essentially on par with (i.e., within 1% of) each other at all core counts. Second, even though the NDP system does not provide any performance improvement for Class 2b functions, it also does not hurt performance. Figure 13 shows the AMAT for our two representative functions. When PLYgemver executes on the host CPU system, up to 77% of the memory latency comes from accessing L3 and DRAM, which can be explained by the function’s medium LFMR (0.5). For SPLLucb, even though up to 73% of memory latency comes from L1 accesses, some requests still hit in the L3 cache (its LFMR is 0.2), translating to around 10% of the memory latency. However, the latency that comes from L3 + DRAM for the host CPU system is similar to the latency to access DRAM in the NDP system, resulting in similar performance between the host CPU system and the NDP system.

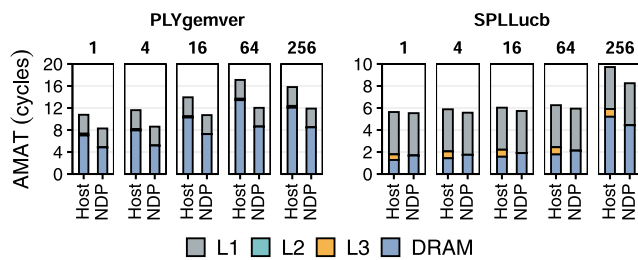


FIGURE 13. AMAT for representative Class 2b functions.

We make a similar observation for the energy consumption for the host CPU system and the NDP system (Figure 14). Even though a small number of memory requests hit in L3, the total energy consumption for both the host CPU system and the NDP system is similar due to L3 and off-chip link energy. For some functions in Class 2b, we observe that the NDP system slightly reduces energy consumption compared to the host CPU system. For example, the NDP system provides an 12% average reduction in energy consumption, across all core counts, compared to the host CPU system for PLYgemver.

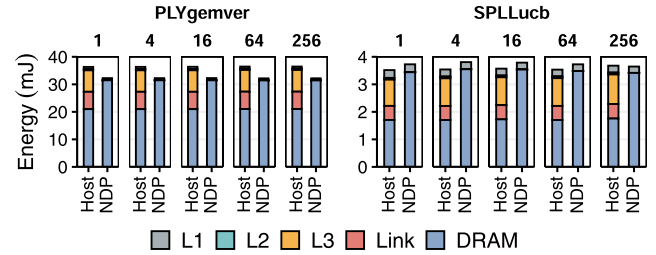


FIGURE 14. Energy breakdown for representative Class 2b functions.

We conclude that while the NDP system does not solve any memory bottlenecks for Class 2b functions, it can be used to reduce the overall SRAM area in the system without any performance or energy penalty (and sometimes with energy savings).

6) CLASS 2c: HIGH TEMPORAL LOCALITY, HIGH AI, LOW LFMR, AND LOW MPKI (COMPUTE-BOUND FUNCTIONS)

Aside from one exception (PLYSymm), all of the 11 functions in this class exhibit high temporal locality. When combined with the high AI and low memory intensity, we find that these characteristics significantly impact how the NDP system performance scales for this class. Figure 5(f) shows the host CPU system and the NDP system performance for HPGSpM and RODNw, two representative functions from the class. We make two observations from the figure. First, the host CPU system performance is always greater than the NDP system performance (by 44% for HPGSpM and 54% for RODNw, on average). The high AI (more than 12 ops per cache line), combined with the high temporal locality and low MPKI, enables these functions to make excellent use of the host CPU system resources. Second, both of the functions benefit greatly from prefetching in the host CPU system. This is a direct result of these functions’ high spatial locality, which allows the prefetcher to be highly accurate and effective in predicting which lines to retrieve from main memory.

Figure 15 shows the energy breakdown consumption for the two representative Class 2c functions. We make two observations. First, the host CPU system is 77% more energy-efficient than the NDP system for HPGSpM, on average across all core counts. Second, the NDP system provides energy savings over the host CPU system at high core counts for RODNw (up to 65% at 256 cores). When the core count increases, the aggregate L1 cache capacity across all cores increases as well, which in turn decreases the number of L1 cache misses. Compared to executing on a single core, executing on 256 cores decreases the L1 cache miss count by 43%, reducing the memory subsystem energy consumption by 40%. However, due to RODNw’s medium LFMR of 0.5, the host CPU system still suffers from L2 and L3 cache misses at high core counts, which require the large L3 and off-chip link energy. In contrast, the NDP system eliminates the energy of accessing the L3 cache and the off-chip link energy by directly sending L1 cache misses to DRAM, which, at high

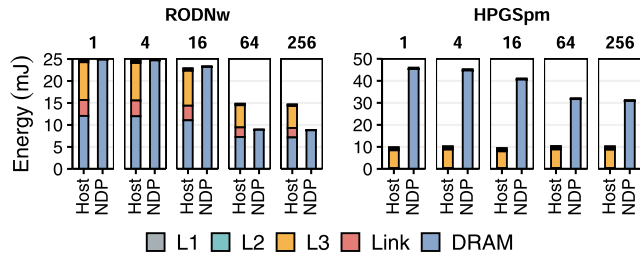


FIGURE 15. Energy breakdown for representative Class 2c functions.

core counts, leads to lower energy consumption than the host CPU system.

We conclude that Class 2c functions do not experience large memory bottlenecks and are not a good fit for the NDP system in terms of performance. However, the NDP system can sometimes provide energy savings for functions that experience medium LFMR.

D. EFFECT OF THE LAST-LEVEL CACHE SIZE

The bottleneck classification we present in Section III-C depends on two key architecture-dependent metrics (LFMR and MPKI) that are directly affected by the parameters and the organization of the cache hierarchy. Our analysis in Section III-C partially evaluates the effect of caching by scaling the aggregated size of the private (L1/L2) caches with the number of cores in the system while maintaining the size of the L3 cache fixed at 8 MB for the host CPU system. However, we also need to understand the impact of the L3 cache size on our bottleneck classification analysis. To this end, this section evaluates the effects on our bottleneck classification analysis of using an alternative cache hierarchy configuration, where we employ a Non-Uniform Cache Architecture (NUCA) [291] model to scale the size of the L3 cache with the number of cores in the host CPU system.

In this configuration, we maintain the sizes of the private L1 and L2 caches (32 kB and 256 kB per core, respectively) while increasing the shared L3 cache size with the core count (we use 2 MB/core) in the host CPU system. The cores, shared L3 caches, and DRAM memory controller are interconnected using a 2D-mesh Network-on-Chip (NoC) [292]–[299] of size $(n + 1) \times (n + 1)$ (an extra interconnection dimension is added to place the DRAM memory controllers). To faithfully simulate the NUCA model (e.g., including network contention in our simulations), we integrate the M/D/1 network model proposed by ZSim++ [300] in our DAMOV simulator [158]. We use a latency of 3 cycles per hop in our analysis, as suggested by prior work [301]. We adapt our energy model to account for the energy consumption of the NoC in the NUCA system. We consider router energy consumption of 63 pJ per request and energy consumed per link traversal of 71 pJ, same as previous work [251].

Figure 16 shows the performance scalability curves for representative functions from each one of our bottleneck classes presented in Section III-C for the baseline host CPU

system (*Host with 8MB Fixed LLC*), the host CPU NUCA system (*Host with NUCA 2MB/Core LLC*), and the NDP system. We make two observations. First, the observations we make for our bottleneck classification (Section III-C) are *not* affected by increasing the L3 cache size for Classes 1a, 1b, 1c, 2b, and 2c. We observe that Class 1a functions benefit from a large L3 cache size (by up to $1.9 \times / 2.3 \times$ for HSNPO/LIGPrkEmd at 256 cores). However, the NDP system still provides performance benefits compared to the host CPU NUCA system. We observe that increasing the L3 size reduces some of the pressure on main memory but cannot fully reduce the DRAM bandwidth bottleneck for Class 1a functions. Functions in Class 1b do *not* benefit from extra L3 capacity (we do not observe a decrease in LFMR or MPKI). Functions in Class 1c do *not* benefit from extra L3 cache capacity. We observe that the private L1 and L2 caches capture most of their data locality, as mentioned in Section III-C3, and thus, these functions do *not* benefit from increasing the L3 size. Functions in Class 2b do *not* benefit from extra L3 cache capacity, which can even lead to a decrease in performance at high core counts for the host CPU NUCA system in some Class 2b functions due to long NUCA L3 access latencies. For example, we observe that PLYgemver’s performance drops 18% when increasing the core count from 64 to 256 in the host CPU NUCA system. We do *not* observe such a performance drop for the host CPU system with fixed LLC size. The performance drop in the host CPU NUCA system is due to the increase in the number of hops that L3 requests need to travel in the NoC at high core counts, which increase the function’s AMAT. Class 2c functions benefit from a larger last-level cache. We observe that their performance improves by $1.3 \times / 1.2 \times$ for HPGSpM/RODNw compared to the host CPU system with 8MB fixed LLC at 256 cores.

Second, we observe two different types of behavior for functions in Class 2a. Since cache conflicts are the major bottleneck for functions in this class, we observe that increasing the L3 cache size can mitigate this bottleneck. In Figure 16, we observe that for both PLYGramSch and SPLFftRev, the host system with NUCA 2MB/Core LLC provides better performance than the host system with 8MB fixed LLC. However, the NDP system can still provide performance benefits in case of contention on the L3 NoC (e.g., in SPLFftRev). For example, the NDP system provides 14% performance improvement for SPLFftRev compared to the NUCA system (with 512 MB L3 cache) for 256 cores.

In summary, we conclude that the key takeaways and observations we present in our bottleneck classification in Section III-C are also valid for a host system with a shared last-level cache whose size scales with core count. In particular, different workload classes get affected by an increase in L3 cache size as expected by their characteristics distilled by our classification.

Figure 17 shows the energy consumption for representative functions from each one of our bottleneck classes presented in Section III-C. We observe that the NDP system can provide

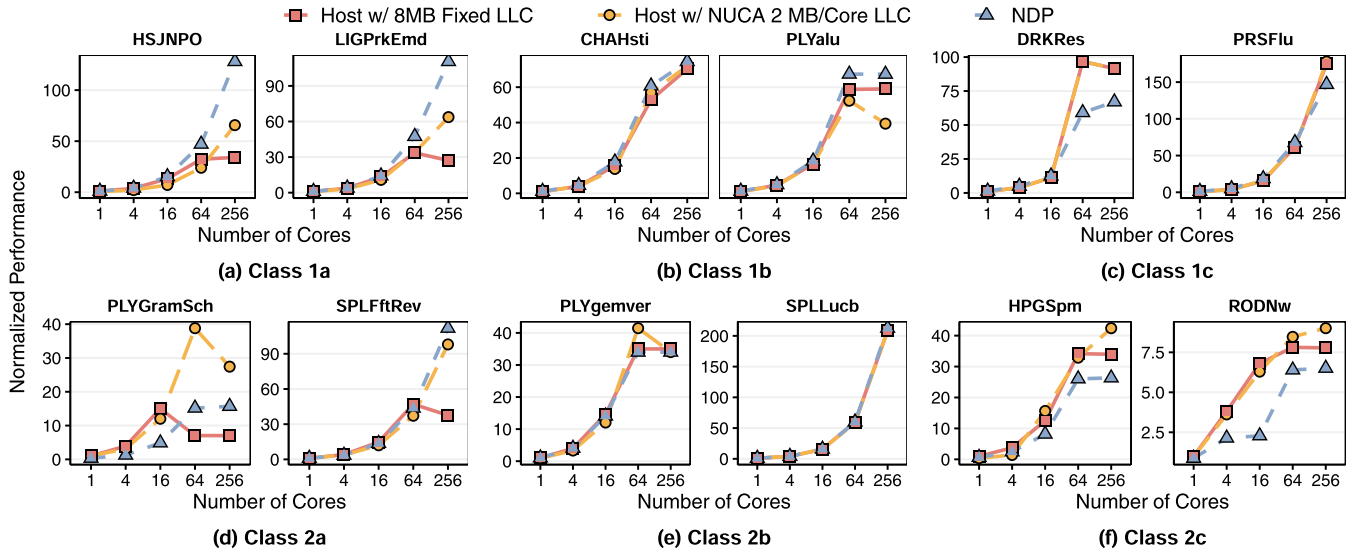


FIGURE 16. Performance of the host and the NDP system as we vary the LLC size, normalized to one host core with a fixed 8MB LLC size.

substantial energy savings for functions in different bottleneck classes, even compared against a system with very large (e.g., 512 MB) cache sizes. We make the following observations for each bottleneck class:

- *Class 1a*: First, for both representative functions in this bottleneck class, the host CPU NUCA system and the NDP system reduce energy consumption compared to the baseline host CPU system. However, we observe that the NDP system provides larger energy savings than the host CPU NUCA system. On average, across all core counts, the NDP system and the host CPU NUCA system reduce energy consumption compared to the host CPU system for *HSJNPO/LIGPrkEmd* by 46%/65% and 25%/22%, respectively. Second, at 256 cores, the host CPU NUCA system provides larger energy savings than the NDP system for both representative functions. This happens because at 256 cores, the large L3 cache (i.e., 512 MB) captures a large portion of the dataset for these functions, reducing costly DRAM traffic. The host CPU NUCA system reduces energy consumption compared to the host CPU system for *HSJNPO/LIGPrkEmd* at 256 cores by $2.0\times/2.2\times$ while the NDP system reduces energy consumption by $1.6\times/1.8\times$. The L3 cache capacity needed to make the host CPU NUCA system more energy efficient than the NDP system is very large (512 MB SRAM), which is likely not cost-effective.
- *Class 1b*: First, for *CHAHsti*, the host CPU NUCA system *increases* energy consumption compared to the host CPU system by 9%, on average across all core counts. In contrast, the NDP system *reduces* energy consumption by 57%. Due to its low spatial and temporal locality (Figure 3), this function does not benefit from a deep cache hierarchy. In the host CPU NUCA

system, the extra energy from the large amount of NoC traffic further increases the cache hierarchy’s overall energy consumption. Second, for *PLYalu*, the host CPU NUCA system and the NDP system reduce energy consumption compared to the host CPU system by 76% and 23%, on average across all core counts. Even though the increase in LLC size does not translate to performance improvements, the large LLC sizes in the host CPU NUCA system aid to reduce DRAM traffic, thereby providing energy savings compared to the baseline host CPU system.

- *Class 1c*: First, for *DRKRes*, the host CPU NUCA system reduces energy consumption compared to the host CPU system by 15%, on average across all core counts. In contrast, the NDP system increases energy consumption by 30%, which is due to the function’s medium LFM (Section III-C3). Second, for *PRSFtu*, we observe that the NDP system provides large energy savings than the host CPU NUCA system. The host CPU NUCA system reduces energy consumption compared to the host CPU system by 21%, while the NDP system reduces energy consumption by 25%, on average across all core counts. However, the energy savings of both host CPU NUCA and NDP systems compared to the host CPU system reduces at high-enough core counts (the energy consumption of the host CPU NUCA system (NDP system) is $0.6\times$ ($0.9\times$) that of the host CPU system at 64 cores and $1.1\times$ ($1.3\times$) that of the host CPU system at 256 cores). This result is expected for Class 1c functions since the functions in this class have decreasing LFM, i.e., the functions effectively utilize the private L1/L2 caches in the host CPU system at high-enough core counts.
- *Class 2a*: First, for *PLYGramSch*, compared to the host CPU system the host CPU NUCA system reduces

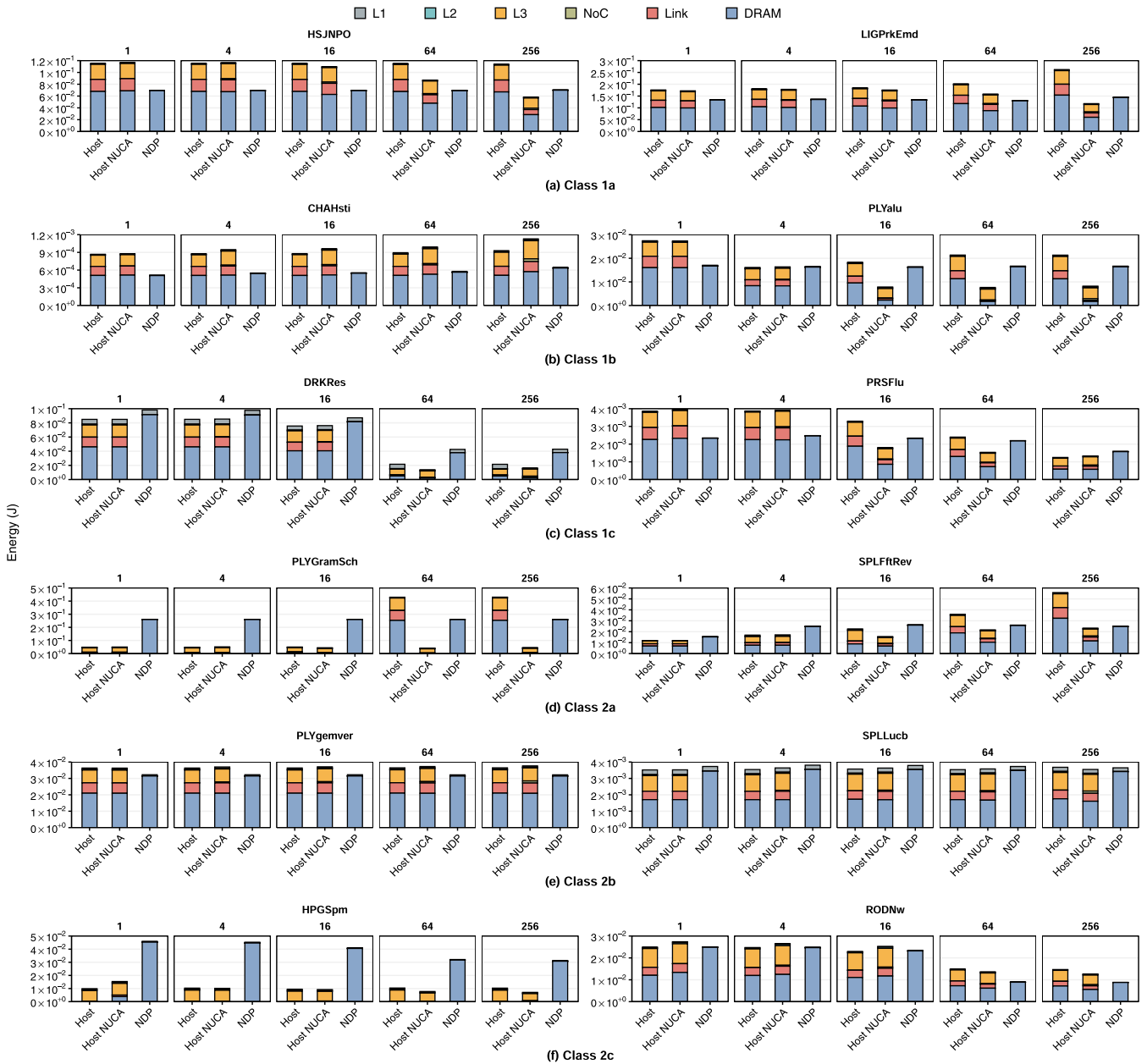


FIGURE 17. Energy of the host and the NDP system as we vary the LLC size. *Host* refers to the host system with a fixed 8MB LLC size; *Host NUCA* refers to the host system with 2MB/Core LLC.

energy consumption by 2.53× and the NDP system increases energy consumption by 55%, on average across all core counts. Even though at high core counts (64 and 256 cores) the host CPU NUCA system provides larger energy savings than the NDP system compared to the host CPU system (the host CPU NUCA system and the NDP system reduce energy consumption compare to the host CPU system by 9× and 65% respectively, averaged across 64 and 256 cores), such large energy savings come at the cost of very large (e.g., 512 MB) cache sizes. Second, for *SPLFftRev*, the host CPU NUCA system and the NDP system reduce energy consumption compared to the host CPU system by 42%

and 7%, on average across all core counts. The NDP system increases energy consumption compared to the host CPU system at low core counts (an increase of 33%, averaged across 1, 4, and 16 cores). However, it provides similar energy savings as the host CPU NUCA system for large core counts (99% and 75% energy reduction compare to the host CPU system for the host CPU NUCA system and the NDP system, respectively, averaged across 64 and 256 cores counts). Since the function suffers from high network contention, the increase in core count increases NoC traffic, which in turn increases energy consumption for the host CPU NUCA system. We conclude that the NDP system provides energy sav-

ings for Class 2a applications compared to the host CPU system at lower cost than the host CPU NUCA system.

- *Class 2b*: First, for `PLYgemver`, the host CPU NUCA system increases energy consumption compared to the host CPU system by 2%, on average across all core counts. In contrast, the NDP system reduces energy consumption by 13%. This function does not benefit from large L3 cache sizes since Class 2b functions are bottlenecked by L1 capacity. Thus, the NoC only adds extra static and dynamic energy consumption. Second, for `SPLLucb`, the host CPU NUCA system consumes the same energy as the host CPU system while the NDP system increases energy consumption by 5%, averaged across all core counts.
- *Class 2c*: For both representative functions in this class, the host CPU NUCA system reduces energy consumption compared to the host CPU system while the NDP system increases energy consumption. For `HPGSPm/RODNw`, the host CPU NUCA system reduces energy consumption by 6%/9% while the NDP system increases energy consumption by 74%/22%, averaged across all core counts. This result is expected since Class 2c functions are compute-bound and highly benefit from a deep cache hierarchy.

In conclusion, the NDP system can provide substantial energy savings for functions in different bottleneck classes, even compared against a system with very large (e.g., 512 MB) cache sizes.

E. VALIDATION AND SUMMARY OF OUR WORKLOAD CHARACTERIZATION METHODOLOGY

In this section, we present the validation and a summary of our new workload characterization methodology. First, we use the remaining 100 memory-bound functions we obtain from *Step 1* (see Section III-A) to validate our workload characterization methodology. To do so, we calculate the accuracy of our workload classification by using the remaining 100 memory-bound functions, which were not used to identify the six classes we found and described in Section III-C. Second, we present a summary of the key metrics we obtain for all 144 memory-bound functions, including our analysis of the host CPU system and the NDP system using two types of cores (in-order and out-of-order).

1) VALIDATION OF OUR WORKLOAD CHARACTERIZATION METHODOLOGY

Our goal is to evaluate the accuracy of our workload characterization methodically on a large set of functions. To this end, we apply *Step 2* and *Step 3* of our memory bottleneck classification methodology (as described in Sections II-C and II-D) to the remaining 100 memory-bound functions we obtain from *Step 1* (in Section III-A). Then, we perform a two-phase validation to calculate the accuracy of our workload characterization.

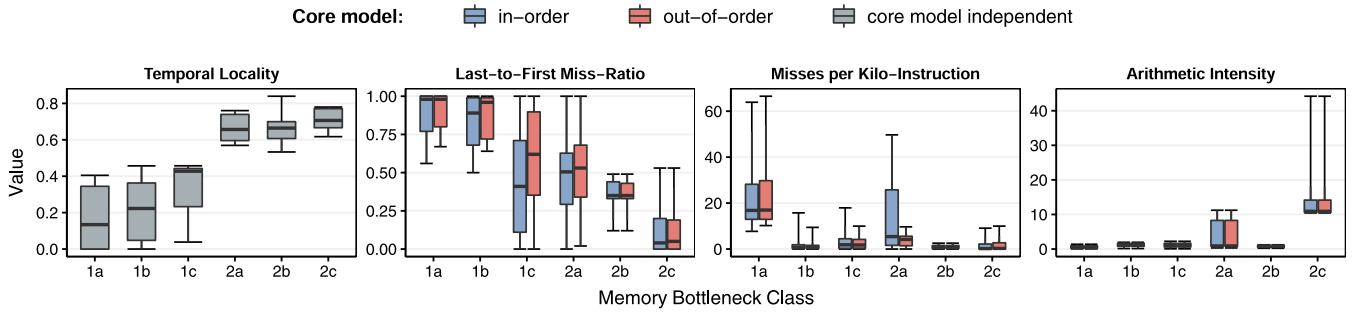
In *phase 1* of our validation, we calculate the threshold values that define the low/high boundaries of each of the four metrics we use to cluster the initial 44 functions in the six memory bottleneck classes in Section III-C (i.e., temporal locality, LFMR, LLC MPKI, and AI). We also include the LFMR curve slope to indicate when the LFMR increases, decreases or stays constant as we scale the core count. We calculate the threshold values for a metric M by computing the middle point between (i) the average value of M across the memory bottleneck classes with *low* values of M and (ii) the average value of M across the memory bottleneck classes with *high* values of M values out of the 44 functions. In *phase 2* of our validation, we calculate the accuracy of our workload characterization by classifying the remaining 100 memory-bound functions using the threshold values obtained from *phase 1* and the LFMR curve slope. After *phase 2*, a function is considered to be *accurately* classified into a correct memory bottleneck class if and only if it (1) fits the definition of the assigned class using the threshold values obtained from *phase 1* and (2) follows the expected performance trends of the assigned class when the function is executed in the host CPU system and the NDP system. For example, a function is correctly classified into Class 1a *if and only if* it (1) displays low temporal locality, low AI, high LFMR, high MPKI and (2) the NDP system outperforms the host CPU system as we scale the core count when executing the function. The final *accuracy of our workload characterization methodology* is calculated by computing the percentage of the functions that are *accurately* classified into one of the six memory bottleneck classes.

First, by applying *phase 1* of our two-phase validation, we obtain that the threshold values are: 0.48 for *temporal locality*, 0.56 for *LFMR*, 11.0 for *MPKI*, and 8.5 for *AI*. Second, by applying *phase 2* of our two-phase validation, we find that we can accurately classify 97% of the 100 memory-bound functions into one of our six memory bottleneck classes (i.e., the accuracy of our workload characterization methodology is 97%). We observe that three functions (*Ligra:ConnectedComponents:compute:rMat*, *Ligra:MaximalIndependentSet:edgeMapDense:USA*, and *SPLASH-2:Oceanncp:relax*) could not be accurately classified into their correct memory bottleneck class (Class 1a). We observe that these functions have LLC MPKI values *lower* than the MPKI threshold expected for Class 1a functions. We expect that the accuracy of our methodology can be further improved by incorporating more workloads into our workload suite and fine-tuning each metric to encompass an even larger set of applications.

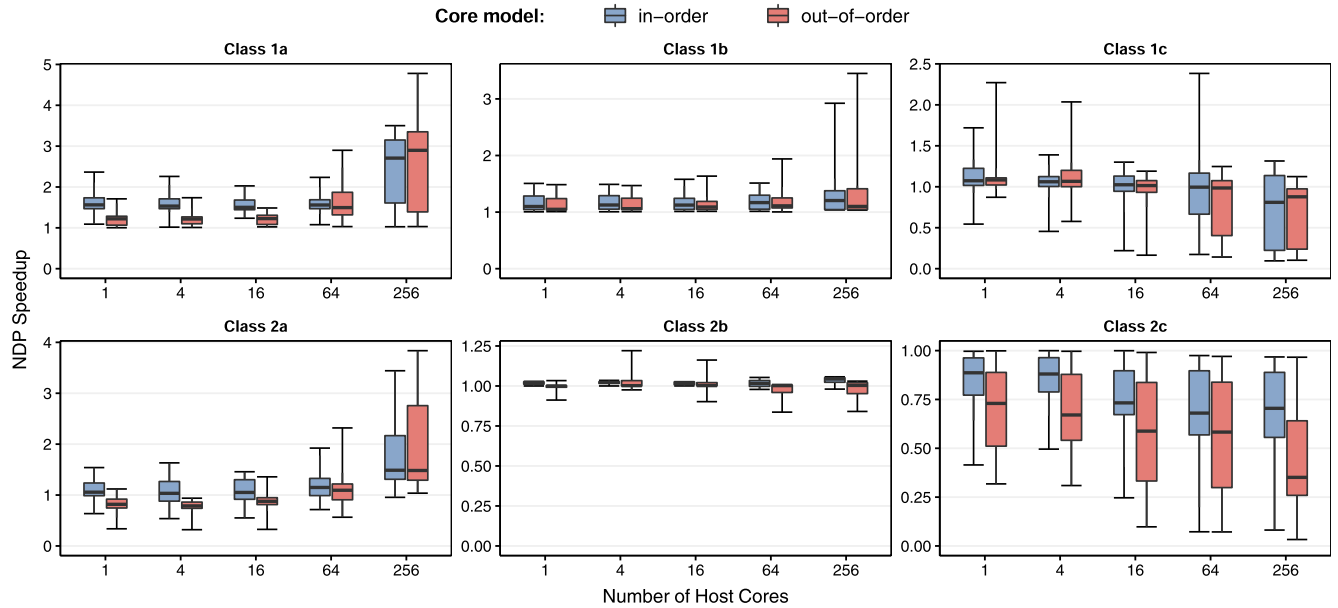
We conclude that our workload characterization methodology can accurately classify a given new application/function into its appropriate memory bottleneck class.

2) SUMMARY OF OUR WORKLOAD CHARACTERIZATION RESULTS

Figure 18a summarizes the metrics we collect for all 144 functions across all core counts (i.e., from 1 to 256 cores)



(a) Summary of the key metrics for each memory bottleneck class.



(b) Summary of NDP speedup for each memory bottleneck class at 1, 4, 16, 64, and 256 cores.

FIGURE 18. Summary of our characterization for all 144 memory-bound functions. Each box is lower-bounded by the first quartile and upper-bounded by the third quartile. The median falls within the box. The inter-quartile range (IQR) is the distance between the first and third quartiles (i.e., box size). Whiskers extend to the minimum and maximum data point values on either sides of the box.

and different core microarchitectures (i.e., out-of-order and in-order cores). The figure shows the distribution of the key metrics we use during our workload characterization for each memory bottleneck class in Section III-C, including architecture-independent metrics (i.e., temporal locality) and architecture-dependent metrics (i.e., AI, LFM, and LLC MPKI). We report the architecture-dependent metrics for two core models: (i) in-order and (ii) out-of-order cores.¹² Together with the out-of-order core model that we use in Section III-C, we incorporate an in-order core model to our analysis, so as to show that our memory bottleneck classification methodology focuses on data movement requirements and works independently of the core microarchitecture. Figure 18b shows the distribution of speedups we observe for

¹²In Section III-C, we collect and report the values of the architecture-independent metrics and architecture-dependent metrics for a subset of 44 representative functions out of the 144 memory-bound functions we identify in *Step 1* of our workload characterization methodology. In Section III-E2, we report values for the *complete set* of 144 memory-bound functions.

when we offload the function to our general-purpose NDP cores, while employing the same core type as the host CPU system.

We make two key observations from Figure 18. First, we observe similar values for each architecture-dependent key metric (i.e., LFM, MPKI, AI) regardless of core type for all 144 functions (in Figure 18a). Second, we observe that the NDP system achieves similar speedups over the host CPU system, when using both in-order and out-of-order core configurations (in Figure 18b). The speedup provided by the NDP system compared to the host CPU system when both systems use out-of-order (in-order) cores for Classes 1a, 1b, 1c, 2a, 2b, and 2c is 1.59 (1.77), 1.22 (1.15), 0.96 (0.95), 1.04 (1.22), 0.94 (1.01), and 0.56 (0.76), respectively, on average across all core counts and functions within a memory bottleneck class. The NDP system greatly outperforms the host CPU system across *all core counts* for Class 1a and 1b functions, with a maximum speedup for the out-of-order (in-order) core model of 4.8 (3.5) and 3.4 (2.9), respectively. The NDP system greatly outperforms the host CPU system at

low core counts for Class 1c functions and at *high core counts* for Class 2a functions, with a maximum speedup for the out-of-order (in-order) core model of 2.3 (2.4) and 3.8 (3.4), respectively. The NDP system provides a modest speedup compared to the host CPU system across *all core counts* for Class 2b functions and slowdown for Class 2c functions, with a maximum speedup for the out-of-order (in-order) core model of 1.2 (1.1) and 1.0 (1.0), respectively. We observe that, averaged across all classes and core types, the average speedup provided by the NDP system using in-order cores is 11% higher than the average speedup offered by the NDP system using out-of-order cores. This is because the host CPU system with out-of-order cores can hide the performance impact of memory access latency to some degree (e.g., using dynamic instruction scheduling) [57], [58], [183], [184], [240], [302]. On the other hand, the host CPU system using in-order cores has little tolerance to hide memory access latency [57], [58], [183], [184], [240], [302].

We conclude that our methodology to classify memory bottlenecks of applications is *robust* and *effective* since we observe similar trends for the six memory bottleneck classes across a large range of (144) functions and two very different core models.

F. LIMITATIONS OF OUR METHODOLOGY

We identify three limitations to our workload characterization methodology. We discuss each limitation next.

1) NDP ARCHITECTURE DESIGN SPACE

Our methodology uses the same type and number of cores in the host CPU and the NDP system configurations for our scalability analysis (Section III-C) because our main goal is to highlight the performance and energy differences between the host CPU system and the NDP system that are caused by data movement. We do not consider practical limitations related to area or thermal dissipation that could affect the type and the maximum number of cores in the NDP system, because our goal is **not** to propose NDP architectures but to characterize data movement and understand the different data movement bottlenecks in modern workloads. Proposing NDP architectures for the workload classes that our methodology identifies as suitable for NDP is a promising topic for future work.

2) FUNCTION-LEVEL ANALYSIS

We choose to conduct our analysis at a function granularity rather than at the application granularity for two major reasons. First, general-purpose NDP architectures are typically leveraged as accelerators to which only *parts* of the application or specific functions are offloaded [1], [47], [48], [54], [59], [64], [65], [83], [86], [89], [92], [98], [100], [102], [133], [192], [193], [303]–[307], rather than the entire application. Functions typically form natural boundaries for parts of algorithms/applications that can potentially be offloaded. Second, it is well-known that applications go through distinct

phases during execution. Each phase may have different characteristics (e.g., a phase might be more compute-bound, while another one might be more memory-bound) and thus fall into different classes in our analysis. A fine-grained analysis at the function level enables us to identify each of those phases and hence, identify more fine-grained opportunities for NDP offloading. However, the main drawback of function-level analysis is that it does not take into account data movement across function boundaries, which affects the performance and energy benefits the NDP system provides over the host CPU system. For example, the NDP system might hurt overall system performance and energy consumption when a large amount of data needs to be continuously moved between a function executing on the NDP cores and another executing on the host CPU cores [63], [74].

3) OVERESTIMATING NDP POTENTIAL

Offloading kernels to NDP cores incurs overheads that our analysis does not account for (e.g., maintaining coherence between the host CPU and the NDP cores [63], [74], efficiently synchronizing computation across NDP cores [101], [140], providing virtual memory support for the NDP system [47], [55], [308], and dynamic offloading support for NDP-friendly functions [48]). Such overheads can impact the performance benefits NDP can provide when considering the end-to-end application. However, deciding how to and whether or not to offload computation to NDP is an open research topic, which involves several architecture-dependent components in the system, such as the following two examples. First, maintaining coherence between the host CPU and the NDP cores is a challenging task that recent works tackle [63], [74]. Second, enabling efficient synchronization across NDP cores is challenging due to the lack of shared caches and hardware cache coherence protocols in NDP systems. Recent works, such as [101], [309], provide solutions to the NDP synchronization problem. Therefore, to focus our analysis on the data movement characteristics of workloads and the broad benefits of NDP, we minimize our assumptions about our target NDP architecture, making our evaluation as broadly applicable as possible.

IV. DAMOV: THE DATA MOVEMENT BENCHMARK SUITE

In this section, we present DAMOV, the DATA MOVEMENT Benchmark Suite. DAMOV is the collection of the 144 functions we use to drive our memory bottleneck classification in Section III. The benchmark suite is divided into each one of the six classes of memory bottlenecks presented in Section III. DAMOV is the first benchmark suite that encompasses *real* applications from a diverse set of application domains tailored to stress different memory bottlenecks in a system. We present the complete description of the functions in DAMOV in Appendix A. We highlight the benchmark diversity of the functions in DAMOV in Section IV-A. We open source DAMOV [158] to facilitate further rigorous

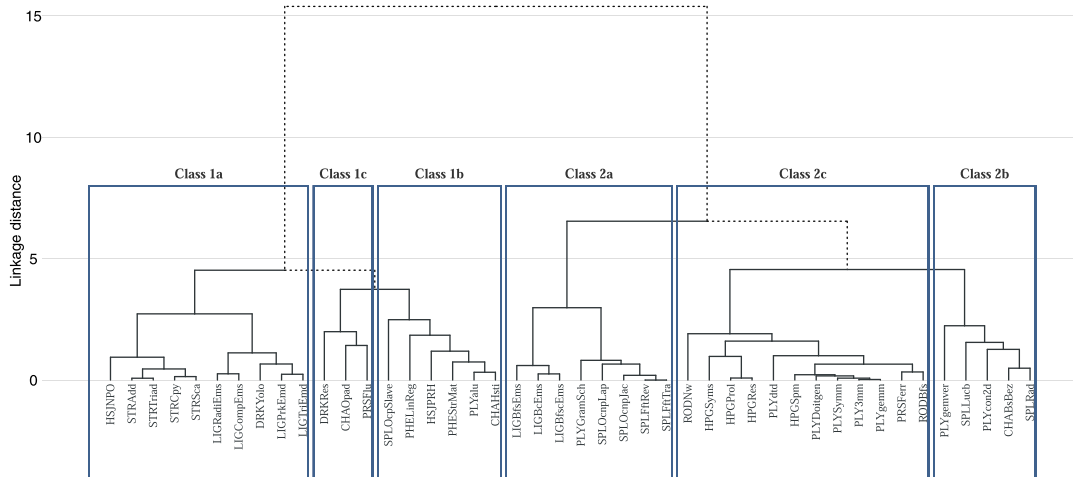


FIGURE 19. Hierarchical clustering of 44 representative functions.

research in mitigating data movement bottlenecks, including in near data processing.

A. BENCHMARK DIVERSITY

We perform a hierarchical clustering algorithm with the 44 representative functions we employ in Section III-C.¹³ Our goal is to showcase our benchmark suite's diversity and observe whether a clustering algorithm produces a noticeable difference from the application clustering presented Section III. The hierarchical clustering algorithm [310] takes as input a dataset containing features that define each object in the dataset. The algorithm works by incrementally grouping objects in the dataset that are similar to each other in terms of some distance metric (called *linkage distance*), which is calculated based on the features' values. Two objects with a short linkage distance have more affinity to each other than two objects with a large linkage distance. To apply the hierarchical clustering algorithm, we create a dataset where each object is one of the 44 representative functions from DAMOV. We use as features the same metrics we use for our analysis, i.e., temporal locality, MPKI, LFMR, and AI. We also include the LFMR curve slope to indicate when the LFMR increases, decreases or stays constant when scaling the core count. We use Euclidean distance [310] to calculate the linkage distance across features in our dataset. We evaluate other linkage distance metrics (such as Manhattan distance [310]), and we observe similar clustering results.

Figure 19 shows the dendrogram that the hierarchical clustering algorithm produces for our 44 representative functions. We indicate in the figure the application class each function belongs to, according to our classification. We make three observations from the figure.

First, our benchmarks exhibit a wide range of behavior diversity, even among those belonging to the same class. For example, we observe that the functions from Class 1a

¹³In Section IV-A, we use the same 44 representative functions that we use during our bottleneck classification instead of the entire set of 144 functions in DAMOV, in order to visualize better the clustering produced by the hierarchical clustering algorithm.

are divided into two groups, with a linkage distance of 3. Intuitively, functions in the first group (HSJNPO, STRAdd, STRCpy, STRSca, STRTriad) have regular access patterns while functions in the second group (DRKYolo, LIGCompEms, LIGPrkEmd, LIGRadiEms) have irregular access patterns. We observe a similar clustering in Section III-C1.

Second, we observe that our application clustering (Section III-C) matches the clustering that the hierarchical clustering algorithm provides (Figure 19). From the dendrogram root, we observe that the right part of the dendrogram consists of functions with high temporal locality (from Classes 2a, 2b, and 2c). Conversely, the left part of the dendrogram consists of functions with low temporal locality (from Classes 1a, 1b, and 1c). The functions in the right and left part of the dendrogram have a high linkage distance (higher than 15), which implies that the metrics we use for our clustering are significantly different from each other for these functions. Third, we observe that functions within the same class are clustered into groups with a linkage distance lower than 5. This grouping matches the six classes of data movement bottlenecks present in DAMOV. Therefore, we conclude that our methodology can successfully cluster functions into distinct classes, each one representing a different memory bottleneck.

We conclude that (i) DAMOV provides a heterogeneous and diverse set of functions to study data movement bottlenecks and (ii) our memory bottleneck clustering methodology matches the clustering provided by a hierarchical clustering algorithm (this section; Figure 19).

V. CASE STUDIES

In this section, we demonstrate how our benchmark suite is useful to study open questions related to NDP system designs. We provide four case studies. The first study analyzes the impact of load balance and communication on NDP execution. The second study assesses the impact of tailored NDP accelerators on our memory bottleneck analysis. The third

study evaluates the effect of different core designs on NDP system performance. The fourth study analyzes the impact of fine-grained offloading (i.e., offloading small blocks of instructions to NDP cores) on performance.

A. CASE STUDY 1: IMPACT OF LOAD BALANCE AND INTER-VULT COMMUNICATION ON NDP SYSTEMS

Communication between NDP cores is one of the key challenges for future NDP system designs, especially for NDP architectures based on 3D-stacked memories, where accessing a remote vault incurs extra latency overhead due to network traffic [46], [101], [311]. This case study aims to evaluate the load imbalance and inter-vault communication that the NDP cores experience when executing functions from the DAMOV benchmark suite. We statically map a function to an NDP core, and we assume that NDP cores are connected using a 6×6 2D-mesh Network-on-Chip (NoC), similar to previous works [66], [70], [312]–[314]. Figure 20 shows the performance overhead that the interconnection network imposes to NDP cores when running several functions from our benchmark suite. We report performance overheads of functions from different bottleneck classes (i.e., from Classes 1a, 1b, 2a, and 2b) that experience at least 5% of performance overhead due to the interconnection network. We calculate the interconnection network performance overhead by comparing performance with the 2D-mesh versus that with an ideal zero-latency interconnection network. We observe that the interconnection network performance overhead varies across functions, with a minimum overhead of 5% for SPLOcpSlave and a maximum overhead of 26% for SPLLucb.

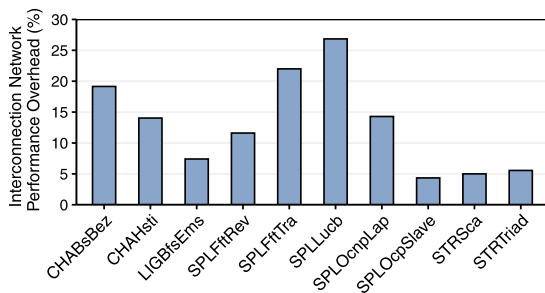


FIGURE 20. Interconnection network performance overhead in our NDP system.

We further characterize the traffic of memory requests injected into the interconnection network for these functions, aiming to understand the communication patterns across NDP cores. Figure 21 shows the distribution of all memory requests (y-axis) in terms of how many hops they need to travel in the NoC between NDP cores (x-axis) for each function. We make the following observations. First, we observe that, on average, 40% of all memory requests need to travel 3 to 4 hops in the NoC, and less than 5% of all requests are issued to a local vault (0 hops). Even though the functions follow different memory access patterns, they all inject similar

network traffic into the NoC.¹⁴ Therefore, we conclude that the NDP design can be further optimized by (i) employing more intelligent data mapping and scheduling mechanisms that can efficiently allocate data nearby the NDP core that accesses the data (thereby reducing inter-vault communication and improving data locality) and (ii) designing interconnection networks that can better fit the traffic patterns that NDP workloads produce. The DAMOV benchmark suite can be used to develop new ideas as well as evaluate existing ideas in both directions.

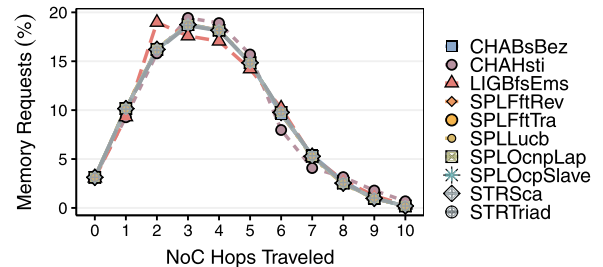


FIGURE 21. Distribution of NoC hops traveled per memory request.

B. CASE STUDY 2: IMPACT OF NDP ACCELERATORS ON OUR MEMORY BOTTLENECK ANALYSIS

In our second case study, we aim to leverage our memory bottleneck classification to evaluate the benefits an NDP accelerator provides compared to the same accelerator accessing memory externally. We use the Aladdin accelerator simulator [315] to tailor an accelerator for an application function. Aladdin works by estimating the performance of a custom accelerator based on the data-flow graph of the application. The main difference between an NDP accelerator and a regular accelerator (i.e., compute-centric accelerator) is that the former is placed in the logic layer of a 3D-stacked memory device and thus can leverage larger memory bandwidth, shorter memory access latency, and lower memory access energy, compared to the compute-centric accelerator that is exemplary of existing compute-centric accelerator designs.

To evaluate the benefits of NDP accelerators, we select three functions from our benchmark suite for this case study: DRKYolo (from Class 1a), PLYalu (from Class 1b), and PLY3mm (from Class 2c). We select these functions and memory bottleneck classes because we expect them to benefit the most (or to show no benefit) from the near-memory placement of an accelerator. According to our memory bottleneck analysis, we expect that the functions we select to (i) benefit from NDP due to its high DRAM bandwidth (Class 1a), (ii) benefit from NDP due to its shorter DRAM access latency (Class 1b), or (iii) do not benefit from NDP in any way (Class 2c).

Figure 22 shows the speedup that the NDP accelerator provides for the different functions compared to the compute-centric accelerator. We make four observations.

¹⁴We use the default HMC data interleaving scheme in our experiments (Table 1).

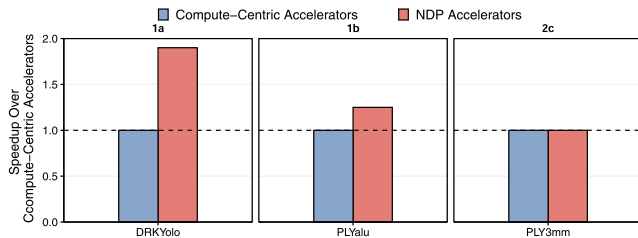


FIGURE 22. Speedup of the NDP accelerators over the compute-centric accelerators for three functions from Classes 1a, 1b, and 2c.

First, as expected based on our classification, the NDP accelerator provides performance benefits compared to the compute-centric accelerator for functions in Classes 1a and 1b. It does not provide performance improvement for the function in Class 2c. Second, the NDP accelerator for DRKYolo shows the largest performance benefits ($1.9\times$ performance improvement compared to the compute-centric accelerator). Since this function is DRAM bandwidth-bound (Class 1a, Section III-C1), the NDP accelerator can leverage the larger memory bandwidth available in the logic layer of the 3D-stacked memory device. Third, we observe that the NDP accelerator also provides speedup ($1.25\times$) for the PLYalu function compared to the compute-centric accelerator, since the NDP accelerator provides shorter memory access latency to the function, which is latency-bound (Class 1b, Section III-C2). Fourth, the NDP accelerator does not provide performance improvement for the PLY3mm function since this function is compute-bound (Class 2c, Section III-C6).

In conclusion, our observations for the performance of NDP accelerators are in line with the characteristics of the three memory bottleneck classes we evaluate in this case study. Therefore, our memory bottleneck classification can be applied to study other types of system configurations, e.g., the accelerators used in this section. However, since NDP accelerators are often employed under restricted area and power constraints (e.g., limited area available in the logic layer of a 3D-stacked memory [63], [74]), the core model of the compute-centric and NDP accelerators cannot always be the same. We leave a thorough analysis that takes area and power constraints in the study of NDP accelerators into consideration for future research.

C. CASE STUDY 3: IMPACT OF DIFFERENT CORE MODELS ON NDP ARCHITECTURES

This case study aims to analyze when a workload can benefit from different core models and numbers of cores while respecting the area and power envelope of the logic layer of a 3D-stacked memory. Many prior works employ 3D-stacked memories as the substrate to implement NDP architectures [1], [46]–[48], [54], [55], [59]–[61], [63]–[70], [74]–[77], [79], [80], [99], [101]–[103], [137], [146], [192], [194], [305], [316]–[324]. However, 3D-stacked memories impose severe area and power restrictions on NDP architec-

tures. For example, the area and power budget of the logic layer of a single HMC vault are 4.4 mm^2 and 312 mW , respectively [1], [63].

In the case study, we perform an iso-area and iso-power performance evaluation of three functions from our benchmark suite. We configure the host CPU system and the NDP system to guarantee an iso-area and iso-power evaluation, considering the area and power budget for a 32-vault HMC device [1], [63]. We use four out-of-order cores with a deep cache hierarchy for the host system configuration and two different NDP configurations: (1) one using six out-of-order NDP cores (*NDP+out-of-order*) and (2) using 128 in-order NDP cores (*NDP+in-order*), without a deep cache hierarchy. We choose functions from Classes 1a, 1b, and 2b for this case study since the major effects distinct microarchitectures have on the memory system are: (a) how much DRAM bandwidth they can sustain, and (b) how much DRAM latency they can hide. Classes 1a, 1b, and 2b are the most affected by memory bandwidth and access latency (as shown in Section III). We choose two representative functions from each of these classes.

Figure 23 shows the speedup provided by the two NDP system configurations compared to the baseline host system. We make two observations. First, in all cases, the *NDP+in-order* system provides higher speedup than the *NDP+out-of-order* system, both compared to the host system. On average across all six functions, the *NDP+in-order* system provides $4\times$ the speedup of the *NDP+out-of-order* system. The larger speedup the *NDP+in-order* system provides is due to the high number of NDP cores in the *NDP+in-order* system. We can fit 128 in-order cores in the logic layer of the 3D-stacked memory as opposed to only six out-of-order cores in the same area/power budget. Second, we observe that the speedup the *NDP+in-order* system provides compared to the *NDP+out-of-order* system does not scale with the number of cores. For example, the *NDP+in-order* system provides *only* $2\times$ the performance of the *NDP+out-of-order* system for DRKYolo and PLYalu, even though the *NDP+in-order* system has $21\times$ the number of NDP cores of the *NDP+out-of-order* system. This implies that even though the functions benefit from a large number of NDP cores available in the *NDP+in-order* system, static instruction scheduling limits performance on the *NDP+in-order* system.

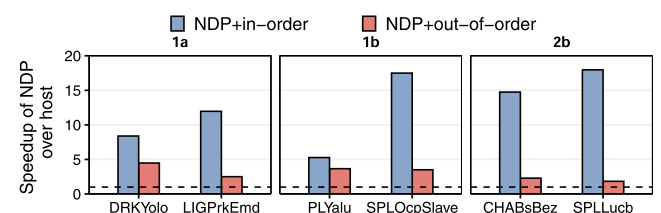


FIGURE 23. Speedup of NDP architectures over 4 out-of-order host CPU cores for two NDP configurations: using 128 in-order NDP cores (*NDP+in-order*) and 6 out-of-order NDP cores (*NDP+out-of-order*) for representative functions from Classes 1a, 1b, and 2b.

We believe, and our previous observations suggest, that an efficient NDP architecture can be achieved by leveraging mechanisms that can exploit both dynamic instruction scheduling and many-core design while fitting in the area and power budget of 3D-stacked memories. For example, past works [57], [58], [183], [184], [224], [325]–[343] propose techniques that enable the benefits of simple and complex cores at the same time, via heterogeneous or adaptive architectures. These ideas can be examined to enable better core and system designs for NDP systems, and DAMOV can facilitate their proper design, exploration, and evaluation.

D. CASE STUDY 4: IMPACT OF FINE-GRAINED OFFLOADING TO NDP ON PERFORMANCE

Several prior works on NDP (e.g., [47], [54], [86], [89], [100], [146], [303], [306], [344]–[346]) propose to identify and offload to the NDP system simple primitives (e.g., instructions, atomic operations). We refer to this NDP offloading scheme as a *fine-grained NDP offloading*, in contrast to a *coarse-grained NDP offloading scheme* that offloads whole functions and applications to NDP systems. A fine-grained NDP offloading scheme provides two main benefits compared to a coarse-grained NDP offloading scheme. First, a fine-grained NDP offloading scheme allows for a reduction in the complexity of the processing elements used as NDP logic, since the NDP logic can consist of simple processing elements (e.g., arithmetic units, fixed function units) instead of entire in-order or out-of-order cores often utilized when employing a coarse-grained NDP offloading scheme. Second, a fine-grained NDP offloading scheme can help developing simple coherence mechanism needed to allow shared host and NDP execution [47]. However, identifying arbitrary NDP instructions can be a daunting task since there is no comprehensive methodology that indicates what types of instructions are good offloading candidates.

As the first step in this direction, we exploit the key insight provided by [151], [347] to identify potential regions of code that can be candidates for fine-grained NDP offloading. [151], [347], [348] show that few instructions are responsible for generating most of the cache misses during program execution in memory-intensive applications. Thus, these instructions are naturally good candidates for fine-grained NDP offloading. Figure 24 shows the distribution of unique basic blocks (x-axis) and the percentage of last-level cache misses (y-axis) the basic block produces for three representative functions from our benchmark suite. We select functions from Classes 1a (LIGKcrEms), 1b (HSJPRH), and 1c (DRKRes) since functions in these classes have higher L3 MPKI than functions in Classes 2a, 2b, and 2c. We observe from the figure that 1% to 10% of the basic blocks in each function are responsible for up to 95.3% of the LLC misses. We call these basic blocks the *hottest* basic blocks.¹⁵ We

¹⁵We observe for the 44 functions we evaluate in Section III that in many cases (for 65% of the evaluated workloads), a single basic block is responsible for 90% to 100% of the LLC misses during the function's execution.

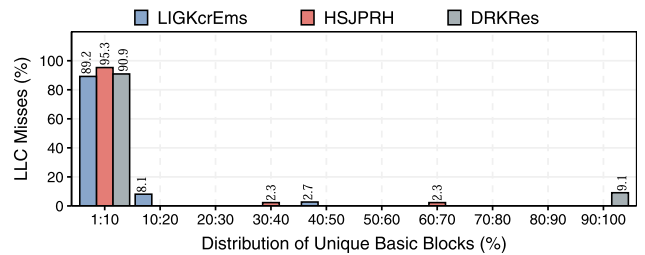


FIGURE 24. Distribution of unique basic blocks (x-axis) and the percentage of last-level cache misses they produce (y-axis) for three representative functions from Classes 1a (LIGKcrEms), 1b (HSJPRH), and 1c (DRKRes).

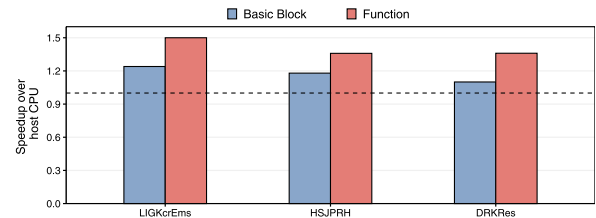


FIGURE 25. Speedup of offloading to NDP the *hottest* basic block in each function versus the entire function.

investigate the data-flow of each basic block and observe that these basic blocks often execute simple read-modify-write operations, with few arithmetic operations. Therefore, we believe that such basic blocks are good candidates for fine-grained offloading. Figure 25 shows the speedup obtained by offloading (i) the hottest basic block we identified for the three representative functions and (ii) the entire function to the NDP system, compared to the host system. Our initial evaluations show that offloading the hottest basic block of each function to the NDP system can provide up to $1.25\times$ speedup compared to the host CPU, which is half of the $1.5\times$ speedup achieved when offloading the entire function. Therefore, we believe that methodically identifying simple NDP instructions can be a promising research direction for future NDP system designs, which our DAMOV Benchmark Suite can help with.

VI. KEY TAKEAWAYS

We summarize the key takeaways from our extensive characterization of 144 functions using our new three-step methodology to identify data movement bottlenecks. We also highlight when NDP is a good architectural choice to mitigate a particular memory bottleneck.

Figure 26 pictorially represents the key takeaways we obtain from our memory bottleneck classification. Based on four key metrics, we classify workloads into six classes of memory bottlenecks. We provide the following key takeaways:

- 1) Applications with low temporal locality, high LFMR, high MPKI, and low AI are DRAM *bandwidth-bound* (Class 1a, Section III-C1). They are bottlenecked by the limited off-chip memory bandwidth as they exert high pressure on main memory. We make three observations for Class 1a applications. First, these

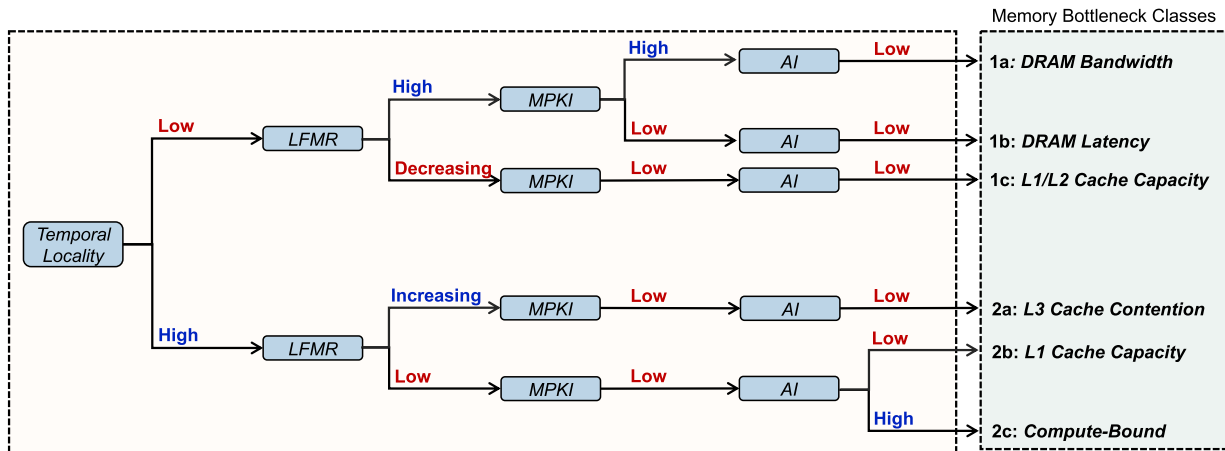


FIGURE 26. Summary of our memory bottleneck classification.

applications do benefit from prefetching since they display a low degree of spatial locality. Second, these applications highly benefit from NDP architectures because they take advantage of the high memory bandwidth available within the memory device. Third, NDP architectures significantly improve energy for these applications since they eliminate the off-chip I/O traffic between the CPU and the main memory.

- 2) Applications with low temporal locality, high LFMR, low MPKI, and low AI are DRAM *latency-bound* (Class 1b, Section III-C2). We make three observations for Class 1b applications. First, these applications do not significantly benefit from prefetching since infrequent memory requests make it difficult for the prefetcher to train successfully on an access pattern. Second, these applications benefit from NDP architectures since they take advantage of NDP's lower memory access latency and the elimination of deep L2/L3 cache hierarchies, which fail to capture data locality for these workloads. Third, NDP architectures significantly improve energy for these applications since they eliminate costly (and unnecessary) L3 cache look-ups and the off-chip I/O traffic between the CPU and the main memory.
- 3) Applications with low temporal locality, decreasing LFMR with core count, low MPKI, and low AI are *bottlenecked by the available L1/L2 cache capacity* (Class 1c, Section III-C3). We make three observations for Class 1c applications. First, these applications are DRAM latency-bound at low core counts, thus taking advantage of NDP architectures, both in terms of performance improvement and energy reduction. Second, NDP's benefits reduce when core count becomes larger, which consequently allows the working sets of such applications to fit inside the cache hierarchy at high core counts. Third, NDP architectures can be a good design choice for such workloads in systems with limited area budget since NDP architectures do not

require large L2/L3 caches to outperform or perform similarly to the host CPU (in terms of both system throughput and energy) for these workloads.

- 4) Applications with high temporal locality, increasing LFMR with core count, low MPKI, and low AI are *bottlenecked by L3 cache contention* (Class 2a, Section III-C4). We make three observations for Class 2a applications. First, these applications benefit from a deep cache hierarchy and do not take advantage of NDP architectures at low core counts. Second, the number of cache conflicts increases when the number of cores in the system increases, leading to more pressure on main memory. We observe that NDP can effectively mitigate such cache contention for these applications without incurring the high area and energy overheads of providing additional cache capacity in the host. Third, NDP can improve energy for these workloads at high core counts, since it eliminates the costly data movement between the last-level cache and the main memory.
- 5) Applications with high temporal locality, low LFMR, low MPKI, and low AI are *bottlenecked by L1 cache capacity* (Class 2b, Section III-C5). We make two observations for Class 2b applications. First, NDP can provide similar performance and energy consumption than the host system by leveraging lower memory access latency and avoiding off-chip energy consumption for these applications. Second, NDP can be used to reduce the overall SRAM area (by eliminating L2/L3 caches) in the system without a performance or energy penalty.
- 6) Applications with high temporal locality, low LFMR, low MPKI, and high AI are *compute-bound* (Class 2c, Section III-C6). We make three observations for Class 2c applications. First, these applications suffer performance and energy penalties due to the lack of a deep L2/L3 cache hierarchy when executed on the NDP architecture. Second, these applications highly

benefit from prefetching due to their high temporal and spatial locality. Third, these applications are not good candidates to execute on NDP architectures.

A. SHAPING FUTURE RESEARCH WITH DAMOV

A key contribution of our work is DAMOV, the first benchmark suite for main memory data movement studies. DAMOV is the collection of 144 functions from 74 different applications, belonging to 16 different benchmark suites or frameworks, classified into six different classes of data movement bottlenecks.

We believe that DAMOV can be used to explore a wide range of research directions on the study of data movement bottlenecks, appropriate mitigation mechanisms, and open research topics on NDP architectures. We highlight DAMOV's usability and potential benefits with four brief case studies, which we summarize below:

- In the first case study (Section V-A), we use DAMOV to evaluate the interconnection network overheads that NDP cores placed in different vaults of a 3D-stacked memory suffer from. We observe that a large portion of the memory requests an NDP core issues go to remote vaults, which increases the memory access latency for the NDP core. We believe that DAMOV can be employed to study better data mapping techniques and interconnection network designs that aim to minimize (i) the number of remote memory accesses the NDP cores execute and (ii) the interconnection network latency overheads.
- In the second case study (Section V-B), we evaluate the benefits that NDP accelerators can provide for three applications from our benchmark suite. We compare the performance improvements an NDP accelerator provides against the compute-centric version of the same accelerator. We observe that the NDP accelerator provides significant performance benefits compared to the compute-centric accelerator for applications in Classes 1a and 1b. At the same time, it does not improve performance for an application in Class 2c. We believe that DAMOV can aid the design of NDP accelerators that target different memory bottlenecks in the system.
- In the third case study (Section V-C), we perform an iso-area/-power performance evaluation to compare NDP systems using in-order and out-of-order cores. We observe that the in-order cores' performance benefits for some applications are limited by the cores' static instruction scheduling mechanism. We believe that better NDP systems can be built by leveraging techniques that enable dynamic instruction scheduling without incurring the large area and power overheads of out-of-order cores. DAMOV can help in the analysis and development of such NDP architectures.
- In the fourth case study (Section V-D), we evaluate the benefits of offloading small portions of code (i.e., a basic

block) to NDP, which simplifies the design of NDP systems. We observe that for many applications, a small percentage of basic blocks is responsible for most of the last-level cache misses. By offloading these basic blocks to an NDP core, we observe a performance improvement of up to $1.25\times$. We believe that DAMOV can be used to identify simple NDP instructions that enable building efficient NDP systems in the future.

VII. RELATED WORK

To our knowledge, this is the first work that methodically characterizes data movement bottlenecks and evaluates the benefits of different data movement mitigation mechanisms, with a focus on Near-Data Processing (NDP), for a broad range of applications. This is also the first work that provides an extensive open-source benchmark suite, with a diverse range of real world applications, tailored to stress different memory-related data movement bottlenecks in a system.

Many past works investigate how to reduce data movement cost using a range of different compute-centric (e.g., prefetchers [56], [189], [244], [349]–[369], speculative execution [57], [58], [183], [184], [349], [370], value-prediction [349], [356], [371]–[387], data compression [388]–[405], approximate computing [40], [371], [406], [407]) and memory-centric techniques [1], [35], [47], [54], [63]–[65], [81], [194], [222], [223], [250], [251], [335], [408]–[417]. These works evaluate the impact of data movement in different systems, including mobile systems [1], [39], [418]–[420], data centers [5], [31], [355], [421]–[425], accelerators-based systems [1], [59], [60], [179], [220], [423], [426], and desktop computers [202], [427], [428]. They use very different profiling frameworks and methodologies to identify the root cause of data movement for a small set of applications. Thus, it is not possible to generalize prior works' findings to other applications than the limited set they analyze.

We highlight two of these prior works, [426] and [1], since they also focus on characterizing applications for NDP architectures. In [426], the authors provide the first work that characterizes workloads for NDP. They analyze five applications (FFT, ray tracing, method of moments, image understanding, data management). The NDP organization [426] targets is similar to [429], where vector processing compute units are integrated into the DDRx memory modules. Even though [426] has a similar goal to our work, it understandably does not provide insights into modern data-intensive applications and NDP architectures as it dates from 2001. Also, [426] focuses its analysis only on a few workloads, whereas we conduct a broader workload analysis starting from 345 applications. Therefore, a new, more comprehensive and rigorous analysis methodology of data movement bottlenecks in modern workloads and modern NDP systems is necessary. A more recent work investigates the memory bottlenecks in widely-used consumer workloads from Google and how NDP can mitigate such bottlenecks [1]. This work focuses its analysis on a small number of consumer work-

loads. Our work presents a comprehensive analysis of a much broader set of applications (345 different applications, and a total of 77K application functions), which allows us to provide a general methodology, a comprehensive workload suite, and general takeaways and guidelines for future NDP research. With our comprehensive analysis, this work is the first to develop a rigorous methodology to classify applications into six groups, which have different characteristics with respect to how they benefit from NDP systems as well as other data movement bottleneck mitigation techniques.

VIII. CONCLUSION

This paper introduces the first rigorous methodology to characterize memory-related data movement bottlenecks in modern workloads and the first data movement benchmark suite, called DAMOV. We perform the first large-scale characterization of applications to develop a three-step workload characterization methodology that introduces and evaluates four key metrics to identify the sources of data movement bottlenecks in real applications. We use our new methodology to classify the primary sources of memory bottlenecks of a broad range of applications into six different classes of memory bottlenecks. We highlight the benefits of our benchmark suite with four case studies, which showcase how representative workloads in DAMOV can be used to explore open-research topics on NDP systems and reach architectural as well as workload-level insights and conclusions. We open-source our benchmark suite and our bottleneck analysis toolchain [158]. We hope that our work enables further studies and research on hardware and software solutions for data movement bottlenecks, including near-data processing.

APPENDIX A

APPLICATION FUNCTIONS IN THE DAMOV BENCHMARK SUITE

We present the list of application functions in each one of the six classes of data movement bottlenecks we identify using our new methodology.

Our benchmark suite is composed of 144 different application functions, collected from 74 different applications. These applications belong to a different set of previously published and widely used benchmark suites. In total, we collect applications from 16 benchmark suites, including: BWA [430], Chai [199], Darknet [215], GASE [208], Hardware Effects [431], Hashjoin [209], HPCC [206], HPCG [207], Ligra [212], PARSEC [202], Parboil [201], PolyBench [432], Phoenix [213], Rodinia [203], SPLASH-2 [205], STREAM [120]. The 144 application functions that are part of DAMOV are listed across six tables, each designating one of the six classes we identify in Section III-C:

- Table 3 lists application functions in Class 1b, i.e., that are DRAM latency-bound (characterized in Section III-C2);
- Table 4 lists application functions in Class 1c, i.e., that are bottlenecked by the available L1/L2 cache capacity (characterized in Section III-C3);
- Table 5 lists application functions in Class 2a, i.e., that are bottlenecked by L3 cache contention (characterized in Section III-C4);
- Table 6 lists application functions in Class 2b, i.e., that are bottlenecked by L1 cache size (characterized in Section III-C5);
- Table 7 lists application functions in Class 2c, i.e., that are compute-bound (characterized in Section III-C6).

In each table we list the benchmark suite, the application name, and the function name. We also list the input size/problem size we use to evaluate each application function.

- Table 2 lists application functions in Class 1a, i.e., that are DRAM bandwidth-bound (characterized in Section III-C1);

TABLE 2. List of application functions in Class 1a.

Class	Suite	Benchmark	Function	Input Set/ Problem Size	Representative Function?
1a	Chai [199]	Transpose	cpu	-m 1024 -n 524288	No
1a	Chai [199]	Vector Pack	run_cpu_threads	-m 268435456 -n 16777216	No
1a	Chai [199]	Vector Unpack	run_cpu_threads	-m 268435456 -n 16777216	No
1a	Darknet [215]	Yolo	gemm	ref	Yes
1a	Hardware Effects [431]	Bandwidth Saturation - Non Temporal	main	ref	No
1a	Hardware Effects [431]	Bandwidth Saturation - Temporal	main	ref	No
1a	Hashjoin [209]	NPO	knuth	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	NPO	ProbeHashTable	-r 12800000 -s 12000000 -x 12345 -y 54321	Yes
1a	Hashjoin [209]	PRH	knuth	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	PRH	lock	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	PRHO	knuth	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	PRHO	radix	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	PRO	knuth	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	PRO	parallel	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	PRO	radix	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Hashjoin [209]	RJ	knuth	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1a	Ligra [212]	Betweenness Centrality	edgeMapSparse	rMat	No
1a	Ligra [212]	Breadth-First Search	edgeMapSparse	rMat	No
1a	Ligra [212]	Connected Components	compute	rMat	No
1a	Ligra [212]	Connected Components	compute	USA	No
1a	Ligra [212]	Connected Components	edgeMapDense	USA	No
1a	Ligra [212]	Connected Components	edgeMapSparse	USA	Yes
1a	Ligra [212]	K-Core Decomposition	compute	rMat	No
1a	Ligra [212]	K-Core Decomposition	compute	USA	No
1a	Ligra [212]	K-Core Decomposition	edgeMapDense	USA	No
1a	Ligra [212]	K-Core Decomposition	edgeMapSparse	rMat	No
1a	Ligra [212]	Maximal Independent Set	compute	rMat	No
1a	Ligra [212]	Maximal Independent Set	compute	USA	No
1a	Ligra [212]	Maximal Independent Set	edgeMapDense	USA	No
1a	Ligra [212]	Maximal Independent Set	edgeMapSparse	rMat	No
1a	Ligra [212]	Maximal Independent Set	edgeMapSparse	USA	No
1a	Ligra [212]	PageRank	compute	rMat	No
1a	Ligra [212]	PageRank	compute	USA	No
1a	Ligra [212]	PageRank	edgeMapDense	USA	Yes
1a	Ligra [212]	Radii	compute	rMat	No
1a	Ligra [212]	Radii	compute	USA	No
1a	Ligra [212]	Radii	edgeMapSparse	USA	No
1a	Ligra [212]	Triangle Count	edgeMapDense	rMat	Yes
1a	SPLASH-2 [205]	Oceanncp	relax	simlarge	No
1a	SPLASH-2 [205]	Oceanncp	relax	simlarge	No
1a	STREAM [120]	Add	Add	50000000	Yes
1a	STREAM [120]	Copy	Copy	50000000	Yes
1a	STREAM [120]	Scale	Scale	50000000	Yes
1a	STREAM [120]	Triad	Triad	50000000	Yes

TABLE 3. List of application functions in Class 1b.

Class	Suite	Benchmark	Function	Input Set/ Problem Size	Representative Function?
1b	Chai [199]	Canny Edge Detection	gaussian	ref	No
1b	Chai [199]	Canny Edge Detection	supression	ref	No
1b	Chai [199]	Histogram - input partition	run_cpu_threads	ref	Yes
1b	Chai [199]	Select	run_cpu_threads	-n 67108864	No
1b	GASE [208]	FastMap	2occ4	Wg2	No
1b	GASE [208]	FastMap	occ4	Wg2	No
1b	Hashjoin [209]	PRH	HistogramJoin	-r 12800000 -s 12000000 -x 12345 -y 54321	Yes
1b	Phoenix [213]	Linear Regression	linear_regression_map	key_file_500MB	No
1b	Phoenix [213]	PCA	main	ref	No
1b	Phoenix [213]	String Match	string_match_map	key_file_500MB	Yes
1b	PolyBench [432]	linear-algebra	lu	LARGE_DATASET	Yes
1b	Rodinia [203]	Kmeans	euclidDist	819200.txt	No
1b	Rodinia [203]	Kmeans	find	819200.txt	No
1b	Rodinia [203]	Kmeans	main	819200.txt	No
1b	Rodinia [203]	Streamcluster	pengain	ref	No
1b	SPLASH-2 [205]	Oceancp	slave2	simlarge	Yes

TABLE 4. List of application functions in Class 1c.

Class	Suite	Benchmark	Function	Input Set/ Problem Size	Representative Function?
1c	BWA [430]	Align	bwa_aln_core	Wg1	No
1c	Chai [199]	Breadth-First Search	comp	USA-road-d	No
1c	Chai [199]	Breadth-First Search	fetch	USA-road-d	No
1c	Chai [199]	Breadth-First Search	load	USA-road-d	No
1c	Chai [199]	Breadth-First Search	run_cpu_threads	USA-road-d	No
1c	Chai [199]	Canny Edge Detection	hystresis	ref	No
1c	Chai [199]	Canny Edge Detection	sobel	ref	No
1c	Chai [199]	Histogram - output partition	run_cpu_threads	ref	No
1c	Chai [199]	Padding	run_cpu_threads	-m 10000 -n 9999	Yes
1c	Chai [199]	Select	fetch	-n 67108864	No
1c	Chai [199]	Stream Compaction	run_cpu_threads	ref	No
1c	Darknet [215]	Resnet	gemm	ref	Yes
1c	Hashjoin [209]	NPO	lock	-r 12800000 -s 12000000 -x 12345 -y 54321	No
1c	Ligra [212]	BFS-Connected Components	edgeMapSparse	rMat	No
1c	Ligra [212]	Triangle Count	compute	rMat	No
1c	Ligra [212]	Triangle Count	compute	USA	No
1c	Ligra [212]	Triangle Count	edgeMapDense	USA	No
1c	PARSEC [202]	Blackscholes	BlkSchlsEqEuroNoDiv	simlarge	No
1c	PARSEC [202]	Fluidaminate	ProcessCollision2MT	simlarge	Yes
1c	PARSEC [202]	Streamcluster	DistL2Float	simlarge	No
1c	Rodinia [203]	Myocyte	find	1000000	No
1c	Rodinia [203]	Myocyte	master	1000000	No

TABLE 5. List of application functions in Class 2a.

Class	Suite	Benchmark	Function	Input Set/ Problem Size	Representative Function?
2a	HPCC [206]	RandomAccess	main	ref	No
2a	HPCC [206]	RandomAccess	update	ref	No
2a	Ligra [212]	Betweenness Centrality	Compute	rMat	No
2a	Ligra [212]	Betweenness Centrality	Compute	USA	No
2a	Ligra [212]	Betweenness Centrality	edgeMapDense	rMat	No
2a	Ligra [212]	Betweenness Centrality	*edgeMapSparse	USA	Yes
2a	Ligra [212]	BFS-Connected Components	Compute	rMat	No
2a	Ligra [212]	BFS-Connected Components	Compute	USA	No
2a	Ligra [212]	BFS-Connected Components	edgeMapSparse	USA	Yes
2a	Ligra [212]	Breadth-First Search	compute	rMat	No
2a	Ligra [212]	Breadth-First Search	compute	USA	No
2a	Ligra [212]	Breadth-First Search	edgeMapDense	rMat	No
2a	Ligra [212]	Breadth-First Search	edgeMapSparse	USA	Yes
2a	Ligra [212]	Connected Components	edgeMapDense	rMat	No
2a	Ligra [212]	Maximal Independent Set	edgeMapDense	rMat	No
2a	Ligra [212]	PageRank	edgeMapDense(Rmat)	rMat	No
2a	Phoenix [213]	WordCount	main	word_100MB	No
2a	PolyBench [432]	linear-algebra	gramschmidt	LARGE_DATASET	Yes
2a	Rodinia [203]	CFD Solver	main	fvcorr.donn.193K	No
2a	SPLASH-2 [205]	FFT2	Reverse	simlarge	Yes
2a	SPLASH-2 [205]	FFT2	Transpose	simlarge	Yes
2a	SPLASH-2 [205]	Oceanncp	jacobcalc	simlarge	No
2a	SPLASH-2 [205]	Oceanncp	laplacalc	simlarge	No
2a	SPLASH-2 [205]	Oceanncp	jacobcalc	simlarge	Yes
2a	SPLASH-2 [205]	Oceanncp	laplacalc	simlarge	Yes
2a	SPLASH-2 [205]	Oceanncp	slave2	simlarge	No

TABLE 6. List of application functions in Class 2b.

Class	Suite	Benchmark	Function	Input Set/ Problem Size	Representative Function?
2b	Chai [199]	Bezier Surface	main_thread	ref	Yes
2b	Hardware Effects [431]	False Sharing - Isolated	main	ref	No
2b	PolyBench [432]	convolution	convolution-2d	LARGE_DATASET	No
2b	PolyBench [432]	linear-algebra	gemver	LARGE_DATASET	Yes
2b	SPLASH-2 [205]	Lucb	Bmod	simlarge	Yes
2b	SPLASH-2 [205]	Radix	slave2	simlarge	Yes

TABLE 7. List of application functions in Class 2c.

Class	Suite	Benchmark	Function	Input Set/ Problem Size	Representative Function?
2c	BWA [430]	Align	bwa_aln_core	Wg2	No
2c	Chai [199]	Transpose	run_cpu_threads	-m 1024 -n 524288	No
2c	Darknet [215]	Alexnet	gemm	ref	No
2c	Darknet [215]	vgg16	gemm	ref	No
2c	Hardware Effects [431]	False Sharing - Shared	main	ref	No
2c	HPCG [207]	HPCG	ComputePrologation	ref	Yes
2c	HPCG [207]	HPCG	ComputeRestriction	ref	Yes
2c	HPCG [207]	HPCG	ComputeSPMV	ref	Yes
2c	HPCG [207]	HPCG	ComputeSYMGS	ref	Yes
2c	Ligra [212]	K-Core Decomposition	edgeMapDense	rMat	No
2c	Ligra [212]	Radii	edgeMapSparse	rMat	No
2c	Parboil [201]	Breadth-First Search	BFS_CPU	ref	No
2c	Parboil [201]	MRI-Gridding	CPU_kernels	ref	No
2c	Parboil [201]	Stencil	cpu_stencil	ref	No
2c	Parboil [201]	Two Point Angular Correlation Function	doCompute	ref	No
2c	PARSEC [202]	Bodytrack	FilterRow	ref	No
2c	PARSEC [202]	Ferret	DistL2Float	ref	Yes
2c	Phoenix [213]	Kmeans	main	ref	No
2c	PolyBench [432]	linear-algebra	3mm	LARGE_DATASET	Yes
2c	PolyBench [432]	linear-algebra	doitgen	LARGE_DATASET	Yes
2c	PolyBench [432]	linear-algebra	gemm	LARGE_DATASET	Yes
2c	PolyBench [432]	linear-algebra	symm	LARGE_DATASET	Yes
2c	PolyBench [432]	stencil	fddt-apml	LARGE_DATASET	Yes
2c	Rodinia [203]	Back Propagation	adjustweights	134217728	No
2c	Rodinia [203]	Back Propagation	layerforward	134217728	No
2c	Rodinia [203]	Breadth-First Search	main	graph1M_6	Yes
2c	Rodinia [203]	Needleman-Wunsch	main	32768	Yes
2c	Rodinia [203]a	Srad	FIN	ref	No
2c	SPLASH-2 [205]	Barnes	computeForces	simlarge	No
2c	SPLASH-2 [205]	Barnes	gravsub	simlarge	No

APPENDIX B REPRESENTATIVE APPLICATION FUNCTIONS

TABLE 8. 44 representative application functions studied in detail in this work.*.

Suite	Benchmark	Function	Short Name	Class	%
Chai [199]	Bezier Surface	Bezier	CHABsBez	2b	100
	Histogram	Histogram	CHAHsti	1b	100
	Padding	Padding	CHAOpad	1c	75.1
Darknet [215]	Resnet 152	gemm_nn	DRKRes	1c	95.2
	Yolo	gemm_nn	DRKYolo	1a	97.1
Hashjoin [209]	NPO	ProbeHashTable	HSJNPO	1a	47.8
	PRH	HistogramJoin	HSJPRH	1b	53.1
HPCG [207]	HPCG	ComputeProlongation	HPGProL	2c	34.3
	HPCG	ComputeRestriction	HPGRes	2c	42.1
	HPCG	ComputeSPMV	HPGSpM	2c	30.5
	HPCG	ComputeSYMGS	HPGSyM	2c	63.6
Ligra [212]	Betweenness Centrality	EdgeMapSparse (USA [218])	LIGBcEms	2a	78.9
	Breadth-First Search	EdgeMapSparse (USA)	LIGBfsEms	2a	67.0
	BFS-Connected Components	EdgeMapSparse (USA)	LIGBfscEms	2a	68.3
	Connected Components	EdgeMapSparse (USA)	LIGCompEms	1a	25.6
	PageRank	EdgeMapDense (USA [218])	LIGPrkEmd	1a	57.2
	Radii	EdgeMapSparse (USA)	LIGRadiEms	1a	67.0
	Triangle	EdgeMapDense (Rmat)	LIGTriEmd	1a	26.7
PARSEC [202]	Ferret	DistL2Float	PRSFerr	2c	18.6
	Fluidanimate	ProcessCollision2MT	PRSFflu	1c	23.9
Phoenix [213]	Linear Regression	linear_regression_map	PHELinReg	1b	76.2
	String Matching	string_match_map	PHEStrMat	1b	38.3
PolyBench [432]	Linear Algebra	3 Matrix Multiplications	PLY3mm	2c	100.0
	Linear Algebra	Multi-resolution analysis kernel	PLYDoitgen	2c	98.3
	Linear Algebra	Matrix-multiply $C=\alpha.A+\beta.B$	PLYgemm	2c	99.7
	Linear Algebra	Vector Mult. and Matrix Addition	PLYgemver	2b	44.4
	Linear Algebra	Gram-Schmidt decomposition	PLYGramSch	2a	100.0
	Linear Algebra	LU decomposition	PLYalu	1b	100.0
	Linear Algebra	Symmetric matrix-multiply	PLYSymm	2c	99.9
	Stencil	2D Convolution	PLYcon2d	2b	100.0
	Stencil	2-D Finite Different Time Domain	PLYdtd	2c	39.8
Rodinia [433]	BFS	BFSGraph	RODBfs	2c	100.0
	Needleman-Wunsch	runTest	RODNw	2c	84.9
SPLASH-2 [205]	FFT	Reverse	SPLFftRev	2a	12.7
	FFT	Transpose	SPLFftTra	2a	8.0
	Lucb	Bmod	SPLLucb	2b	77.6
	Oceanncp	jacobcalc	SPLOcnpJac	2a	30.7
	Oceanncp	laplacecalc	SPLOcnpLap	2a	23.4
	Oceanncp	slave2	SPLOcpSlave	1b	24.4
	Radix	slave_sort	SPLRad	2b	41.1
	Add	Add	STRAdd	1a	98.4
STREAM [120]	Copy	Copy	STRCpy	1a	98.3
	Scale	Scale	STRSca	1a	97.5
	Triad	Triad	STRTriad	1a	99.1

* Short names are encoded as XXXYyyZzz, where XXX is the source application suite, Yyy is the application name, and Zzz is the function (if more than one per benchmark). For graph processing applications from Ligra, we test two different input graphs, so we append the graph name to the short benchmark name as well. The % column indicates the percentage of clock cycles that the function consumes as a fraction of the execution time of the entire benchmark.

APPENDIX C COMPLETE LIST OF EVALUATED APPLICATIONS

TABLE 9. List of the evaluated 345 applications.

Benchmark Suite	Application	Benchmark Suite	Application	Benchmark Suite	Application
ArtraCFD [434]	ArtraCFD		Global Dot Product		disparity
blasr [436]	Long read aligner		Multigrid preconditione		localization
BWA [437]	aln	HPCG [207]	Sparse Matrix Vector Multiplication (SpMV)		mser
	fastmap		Symmetric Gauss-Seidel smoother (SymGS)	SD-VBS - Vision [435]	multi_ncut
	BFS		Vector Update		pca
	BS	IMPICA Workloads [55]	btree		sift
	CEDD		hashtable		stitch
	HSTI		llbenchmark		svm
	HSTO	libvpx [438]	VP8		texture_synthesis
Chai [199]	OOPPAD		VP9		tracking
	OOPTRNS		BC	sort-merge-joins [439]	m-pass
	SC		BellmanFord		m-way
	SELECT		BFS		400.perlbench
	TRNS		BFS-Bitvector		401.bzj2
	VPACK		BFS-CC		403.gcc
	VUPACK		CF		410.bwaves
clstm [441]	clstm	Ligra [212]	Components		416.gamess
	BetwCent		KCore		429.mcf
	BipartiteMatchings		MIS		433.milc
	CC		PageRank		434.zeusmp
	DirOptBFS		PageRankDelta		435.gromacs
CombBLAS [442]	FilteredBFS		Radii		436.cactusADM
	FilteredMIS	Metagraph [443]	Triangle		437.leslie3d
	MCL3D		annotate		444.namd
	Ordering/RCM		classify		445.gobmk
	TopDownBFS		ASUM		447.dealII
	AMG2013		AXPY	SPEC CPU2006 [440]	450.soplex
	CAM-SE	MKL [444]	DOT		453.povray
	Graph500		GEMM		454.calculix
	HACC		GEMV		456.hmmer
	Hash		mri-q		458.sjeng
	homme1_3_6		BFS		459.GemsFDTD
	Integer Sort		cutcp		462.libquantum
	KMI	Parboil [201]	histo		464.h264ref
	LSMS		lbm		465.tonto
	LULESH		mri-gridding		470.lbm
	MCB		sad		471.omnetpp
	miniFE		sgemm		473.astar
	Nekbone		spmv		481.wrf
	QBOX		stencil		482.sphinx3
	SNAP		tpacf		483.xalancbmk
	SPECint2006 "peak"		blackscholes		500.perlbench_r
	UMT2013		bodytrack		502.gcc_r
	AlexNet		canneal		503.bwaves_r
	Darknet19	PARSEC [202]	dedup		505.mcf_r
	Darknet53		facesim		507.cactuBSSN_r
	Densenet 201		ferret		508.namd_r
	Extraction		fluidanimate		510.parest_r
	Resnet 101		fregmine		511.povray_r
	Resnet 152		raytrace		519.lbm_r
	Resnet 18		streamcluster		520.omnetpp_r
	Resnet 34		swaptions		521.wrf_r
	Resnet 50		vips		523.xalancbmk_r
	ResNeXt 101		x264		525.x264_r
	ResNext 152		histogram		526.blender_r
	ResNeXt50		kmeans		527.cam4_r
	VGG-16	Phoenix [213]	linear-regression		531.deepsjeng_r
	Yolo		matrix multiply		538.imagick_r
DBF-5 [446]	TPC-E		pca		541.leela_r
	TPCC DL_DETECT		string_match		544.nab_r
	TPCC HEKATON		word_count		548.exchange2_r
	TPCC NO_WAIT		2mm		549.totoni3d_r
	TPCC SILO		3mm		554.roms_r
	TPCC TIC/OC		atax		557.xz_r
	YCSB DL_DETECT		bicg		600.perlbench_s
	YCSB HEKATON		cholesky		602.gcc_s
	YCSB NO_WAIT		convolution-2d		603.bwaves_s
	YCSB SILO		convolution-3d		605.mcf_s
	YCSB TIC/OC		correlation		607.cactuBSSN_s
	RMI-large [133]		covariance		619.lbm_s
	RMI-small [133]		doitgen		620.omnetpp_s
	RM2-large [133]	PolyBench [432]	durbin		621.wrf_s
	RM2-small [133]		fdtd-apm		623.xalancbmk_s
GASE [208]	FastMap		gemm		625.x264_s
	gale_aln		gemver		627.cam4_s
	BFS		gramschmidt		628.pop2_s
	DeltaStepping		gramschmidt		631.deepsjeng_s
	Incremental PageRank		lu		638.imagick_s
	LDA		lu		641.leela_s
	PageRank		mvt		644.nab_s
	SDG		symp		648.exchange2_s
	SSSP		syz2k		649.totoni3d_s
	Topological Sort		syrk		654.roms_s
	Triangle Counting		trmm		657.xz_s

TABLE 9. (Continued.) List of the evaluated 345 applications.

Hardware Effects [431]	4k aliasing	resectionvolume [449]	resectionvolume	SPLASH-2 [205]	barnes
	bandwidth saturation non-temporal		b+tree		cholesky
	bandwidth saturation temporal		backprop		fft
	branch misprediction sort		bfs		fmm
	branch misprediction unsort		cfid		lu_cb
	branch target misprediction		heartwall		lu_ncb
	cache conflicts		hotspot		ocean_cp
	cache/memory hierarchy bandwidth		hotspot3D		ocean_ncp
	data dependencies		kmeans		radiosity
	denormal floating point numbers		lavaMD		radix
	denormal floating point numbers flush		leukocyte		raytrace
	DRAM refresh interval		lud		volrend
	false sharing		mummergepu		water_nsquared
	hardware prefetching		myocyte		water_spatial
	hardware prefetching shuffle		nn		img-dnn
	hardware store elimination		nw		masstree
	memory-bound program		particlefilter		moses
	misaligned accesses		pathfinder		shore
	non-temporal stores		srad		silo
	software prefetching		streamcluster		specjbb
	store buffer capacity		lda		sphinx
	write combining		libl		xpian
	Hashjoin [209]		NPO		SD-VBS- Cortex [435]
PRH		pca	echo		
PRHO		rbm	exim		
PRO		sphinx	hashmap		
RJ		srr	memcached		
HPCC [206]	RandomAccesses	svd	WHISPER [451]	nfs	
				redis	
				sql	
				tpcc	
				vacation	
			ycsb		
			ZipML [452]	SGD	
			Stream [120]	STREAM	

ACKNOWLEDGMENT

The authors would like to thank the SAFARI Research Group members for valuable feedback and the stimulating intellectual environment they provide. An earlier version of this work was posted on arxiv.org (<https://arxiv.org/pdf/2105.03725.pdf>) on May 8, 2021. Talk videos for this work are available on YouTube, including a short talk video (<https://youtu.be/HkMYuYMuZOg>), a long talk video (<https://youtu.be/GWideVyo0nM>), and a tutorial on the DAMOV framework and benchmarks (<https://youtu.be/GWideVyo0nM?t=8028>).

REFERENCES

[1] A. Boroumand, S. Ghose, Y. Kim, R. Ausavarungnirun, E. Shiu, R. Thakur, D. Kim, A. Kuusela, A. Knies, P. Ranganathan, and O. Mutlu, "Google workloads for consumer devices: Mitigating data movement bottlenecks," in *Proc. ASPLOS*, 2018, pp. 316–331.

[2] O. Mutlu, "Memory scaling: A systems architecture perspective," in *Proc. 5th IEEE Int. Memory Workshop*, May 2013, pp. 21–25.

[3] U. Kang, H.-S. Yu, C. Park, H. Zheng, J. Halbert, K. Bains, S. Jang, and J. S. Choi, "Co-architecting controllers and DRAM to enhance DRAM process scaling," in *Proc. Memory Forum*, vol. 14, 2014.

[4] S. Hong, "Memory technology trend and future challenges," in *Proc. Int. Electron Devices Meeting*, Dec. 2010, pp. 12.4.1–12.4.4.

[5] S. Kanev, J. P. Darago, K. Hazelwood, P. Ranganathan, T. Moseley, G.-Y. Wei, and D. Brooks, "Profiling a warehouse-scale computer," in *Proc. 42nd Annu. Int. Symp. Comput. Archit.*, Jun. 2015, pp. 158–169.

[6] O. Mutlu and L. Subramanian, "Research problems and opportunities in memory systems," *Supercomput. Frontiers Innov.*, vol. 1, no. 3, pp. 19–55, 2014.

[7] O. Mutlu, "Main memory scaling: Challenges and solution directions," in *More Than Moore Technologies for Next Generation Computer Design*. New York, NY, USA: Springer-Verlag, 2015.

[8] J. S. Kim, M. Patel, A. G. Yağlıkçı, H. Hassan, R. Azizi, L. Orosa, and O. Mutlu, "Revisiting RowHammer: An experimental analysis of modern DRAM devices and mitigation techniques," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May 2020, pp. 638–651.

[9] Y. Kim, R. Daly, J. Kim, C. Fallin, J. H. Lee, D. Lee, C. Wilkerson, K. Lai, and O. Mutlu, "Flipping bits in memory without accessing them: An experimental study of DRAM disturbance errors," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 695–711.

[10] O. Mutlu, "The RowHammer problem and other issues we may face as memory becomes denser," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2017, pp. 1116–1121.

[11] S. Ghose, A. G. Yağlıkçı, R. Gupta, D. Lee, K. Kudrolli, W. X. Liu, H. Hassan, K. K. Chang, N. Chatterjee, A. Agrawal, M. O'Connor, and O. Mutlu, "What your DRAM power models are not telling you: Lessons from a detailed experimental study," in *Proc. SIGMETRICS*, Jun. 2018, pp. 1–41.

[12] O. Mutlu and J. S. Kim, "RowHammer: A retrospective," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 8, pp. 1555–1571, May 2020.

[13] P. Frigo, E. Vannacc, H. Hassan, V. V. der Veen, O. Mutlu, C. Giuffrida, H. Bos, and K. Razavi, "TRRespass: Exploiting the many sides of target row refresh," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 747–762.

[14] J. Liu, B. Jaiyen, Y. Kim, C. Wilkerson, and O. Mutlu, "An experimental study of data retention behavior in modern DRAM devices: Implications for retention time profiling mechanisms," in *Proc. ISCA*, Jun. 2013, pp. 60–71.

[15] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, "RAIDR: Retention-aware intelligent DRAM refresh," in *Proc. 39th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2012, pp. 1–12.

- [16] M. Patel, J. S. Kim, and O. Mutlu, "The reach profiler (REAPER): Enabling the mitigation of DRAM retention failures via profiling at aggressive conditions," in *Proc. 44th Annu. Int. Symp. Comput. Archit.*, Jun. 2017, pp. 255–268.
- [17] M. K. Qureshi, D.-H. Kim, S. Khan, P. J. Nair, and O. Mutlu, "AVATAR: A variable-retention-time (VRT) aware refresh for DRAM systems," in *Proc. 45th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, Jun. 2015, pp. 427–437.
- [18] J. A. Mandelman, R. H. Dennard, G. B. Bronner, J. K. DeBrosse, R. Divakaruni, Y. Li, and C. J. Radens, "Challenges and future directions for the scaling of dynamic random-access memory (DRAM)," *IBM J. Res. Develop.*, vol. 46, pp. 187–212, Mar. 2002.
- [19] S. Khan, D. Lee, Y. Kim, A. R. Alameldeen, C. Wilkerson, and O. Mutlu, "The efficacy of error mitigation techniques for DRAM retention failures: A comparative experimental study," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst. (SIGMETRICS)*, 2014, pp. 519–532.
- [20] S. Khan, D. Lee, and O. Mutlu, "PARBOR: An efficient system-level technique to detect data-dependent failures in DRAM," in *Proc. 46th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2016, pp. 239–250.
- [21] S. Khan, C. Wilkerson, Z. Wang, A. R. Alameldeen, D. Lee, and O. Mutlu, "Detecting and mitigating data-dependent DRAM failures by exploiting current memory content," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2017, pp. 27–40.
- [22] D. Lee, Y. Kim, G. Pekhimenko, S. Khan, V. Seshadri, K. Chang, and O. Mutlu, "Adaptive-latency DRAM: Optimizing DRAM timing for the common-case," in *Proc. IEEE 21st Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2015, pp. 489–501.
- [23] D. Lee, S. Khan, L. Subramanian, S. Ghose, R. Ausavarungnirun, G. Pekhimenko, V. Seshadri, and O. Mutlu, "Design-induced latency variation in modern DRAM chips: Characterization, analysis, and latency reduction mechanisms," in *Proc. SIGMETRICS*, 2017, pp. 1–36.
- [24] K. K. Chang, A. G. Yağlıkcı, S. Ghose, A. Agrawal, N. Chatterjee, A. Kashyap, D. Lee, M. O'Connor, H. Hassan, and O. Mutlu, "Understanding reduced-voltage operation in modern DRAM devices: Experimental characterization, analysis, and mechanisms," in *Proc. Int. Conf. Meas. Model. Comput. Syst. (ACM SIGMETRICS)*, Jun. 2017, pp. 1–42.
- [25] K. K. Chang, A. Kashyap, H. Hassan, S. Ghose, K. Hsieh, D. Lee, T. Li, G. Pekhimenko, S. Khan, and O. Mutlu, "Understanding latency variation in modern DRAM chips: Experimental characterization, analysis, and optimization," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Modeling Comput. Sci.*, Jun. 2016, pp. 323–336.
- [26] K. K.-W. Chang, D. Lee, Z. Chishti, A. R. Alameldeen, C. Wilkerson, Y. Kim, and O. Mutlu, "Improving DRAM performance by parallelizing refreshes with accesses," in *Proc. IEEE 20th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2014, pp. 356–367.
- [27] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "Revisiting memory errors in large-scale production data centers: Analysis and modeling of new trends from the field," in *Proc. 45th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, Jun. 2015, pp. 415–426.
- [28] H. David, C. Fallin, E. Gorbatov, U. R. Hanenbutte, and O. Mutlu, "Memory power management via dynamic voltage/frequency scaling," in *Proc. 8th ACM Int. Conf. Autonomic Comput. (ICAC)*, 2011, pp. 31–40.
- [29] Q. Deng, D. Meisner, L. Ramos, T. F. Wenisch, and R. Bianchini, "MemScale: Active low-power modes for main memory," in *Proc. ASPLOS*, 2011, pp. 225–238.
- [30] J. Dean and L. A. Barroso, "The tail at scale," in *Proc. CACM*, 2013, pp. 74–80.
- [31] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafae, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi, "Clearing the clouds: A study of emerging scale-out workloads on modern hardware," in *Proc. ASPLOS*, 2012, pp. 37–48.
- [32] L. Wang, J. Zhan, C. Luo, Y. Zhu, Q. Yang, Y. He, W. Gao, Z. Jia, Y. Shi, S. Zhang, C. Zheng, G. Lu, K. Zhan, X. Li, and B. Qiu, "Big-DataBench: A big data benchmark suite from internet services," in *Proc. IEEE 20th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2014, pp. 488–499.
- [33] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, "Enabling practical processing in and near memory for data-intensive computing," in *Proc. 56th Annu. Design Autom. Conf.*, Jun. 2019, pp. 1–4.
- [34] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, "Processing data where it makes sense: Enabling in-memory computation," in *Proc. MicPro*, 2019, pp. 28–41.
- [35] O. Mutlu, "Intelligent architectures for intelligent machines," in *Proc. Int. Symp. VLSI Design, Autom. Test (VLSI-DAT)*, Aug. 2020, pp. 1–4.
- [36] S. Ghose, A. Boroumand, J. S. Kim, J. Gómez-Luna, and O. Mutlu, "Processing-in-memory: A workload-driven perspective," in *Proc. IBM JRD*, 2019, pp. 1–3.
- [37] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, "A modern primer on processing in memory," in *Emerging Computing: From Devices to Systems—Looking Beyond Moore and Von Neumann*. New York, NY, USA: Springer, 2021.
- [38] S. Wang and E. Ipek, "Reducing data movement energy via online data clustering and encoding," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–13.
- [39] D. Pandiyan and C.-J. Wu, "Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2014, pp. 171–180.
- [40] S. Koppula, L. Orosa, A. G. Yağlıkcı, R. Azizi, T. Shahroodi, K. Kanellopoulos, and O. Mutlu, "EDEN: Enabling energy-efficient, high-performance deep neural network inference using approximate DRAM," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 166–181.
- [41] S. A. McKee, "Reflections on the memory wall," in *Proc. 1st Conf. Comput. Frontiers (CF)*, 2004, p. 162.
- [42] M. V. Wilkes, "The memory gap and the future of high performance memories," *ACM SIGARCH Comput. Archit. News*, vol. 29, no. 1, pp. 2–7, 2001.
- [43] Y. Kim, V. Seshadri, D. Lee, J. Liu, and O. Mutlu, "A case for exploiting subarray-level parallelism (SALP) in DRAM," in *Proc. 39th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2012, pp. 368–379.
- [44] W. A. Wulf and S. A. McKee, "Hitting the memory wall: Implications of the obvious," *ACM SIGARCH Comput. Archit. News*, vol. 23, no. 1, pp. 20–24, 1995.
- [45] S. Ghose, T. Li, N. Hajinazar, D. Senol Cali, and O. Mutlu, "Demystifying complex workload-DRAM interactions: An experimental study," in *Proc. Abstr. SIGMETRICS/Perform. Joint Int. Conf. Meas. Modeling Comput. Syst.*, Jun. 2019, pp. 1–50.
- [46] J. Ahn, S. Hong, S. Yoo, O. Mutlu, and K. Choi, "A scalable processing-in-memory accelerator for parallel graph processing," in *Proc. 42nd Annu. Int. Symp. Comput. Archit.*, Jun. 2015, pp. 105–117.
- [47] J. Ahn, S. Yoo, O. Mutlu, and K. Choi, "PIM-enabled instructions: A low-overhead, locality-aware processing-in-memory architecture," in *Proc. 42nd Annu. Int. Symp. Comput. Archit.*, Jun. 2015, pp. 336–348.
- [48] K. Hsieh, E. Ebrahim, G. Kim, N. Chatterjee, M. O'Connor, N. Vijaykumar, O. Mutlu, and S. W. Keckler, "Transparent offloading and mapping (TOM): Enabling programmer-transparent near-data processing in GPU systems," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 204–216.
- [49] Y. Wang, L. Orosa, X. Peng, Y. Guo, S. Ghose, M. Patel, J. S. Kim, J. G. Luna, M. Sadrosadati, N. M. Ghiasi, and O. Mutlu, "FIGARO: Improving system performance via fine-grained in-DRAM data relocation and caching," in *Proc. 53rd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2020, pp. 313–328.
- [50] Z. Jia, J. Zhan, L. Wang, C. Luo, W. Gao, Y. Jin, R. Han, and L. Zhang, "Understanding big data analytics workloads on modern processors," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 6, pp. 1797–1810, Nov. 2016.
- [51] P.-A. Tsai, C. Chen, and D. Sanchez, "Adaptive scheduling for systems with asymmetric memory hierarchies," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2018, pp. 641–654.
- [52] R. Sites, "It's the memory, stupid!" *MPR*, vol. 10, no. 10, pp. 2–3, Aug. 1996.
- [53] V. Seshadri, T. Mullins, A. Boroumand, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "Gather-scatter DRAM: In-DRAM address translation to improve the spatial locality of non-unit strided accesses," in *Proc. 48th Int. Symp. Microarchitecture*, Dec. 2015, pp. 267–280.
- [54] L. Nai, R. Hadidi, J. Sim, H. Kim, P. Kumar, and H. Kim, "GraphPIM: Enabling instruction-level PIM offloading in graph computing frameworks," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 457–468.
- [55] K. Hsieh, S. Khan, N. Vijaykumar, K. K. Chang, A. Boroumand, S. Ghose, and O. Mutlu, "Accelerating pointer chasing in 3D-stacked memory: Challenges, mechanisms, evaluation," in *Proc. IEEE 34th Int. Conf. Comput. Design (ICCD)*, Oct. 2016, pp. 25–32.
- [56] E. Ebrahimi, O. Mutlu, and Y. N. Patt, "Techniques for bandwidth-efficient prefetching of linked data structures in hybrid prefetching systems," in *Proc. IEEE 15th Int. Symp. High Perform. Comput. Archit.*, Feb. 2009, pp. 7–17.

- [57] O. Mutlu, J. Stark, C. Wilkerson, and Y. N. Patt, "Runahead execution: An alternative to very large instruction windows for out-of-order processors," in *Proc. 9th Int. Symp. High-Perform. Comput. Archit. (HPCA)*, 2003, pp. 129–140.
- [58] M. Hashemi, O. Mutlu, and Y. N. Patt, "Continuous runahead: Transparent hardware acceleration for memory intensive workloads," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–12.
- [59] J. S. Kim, D. Senol Cali, H. Xin, D. Lee, S. Ghose, M. Alser, H. Hassan, O. Ergin, C. Alkan, and O. Mutlu, "GRIM-filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies," *BMC Genomics*, vol. 19, no. S2, pp. 23–40, May 2018.
- [60] D. S. Cali, G. S. Kalsi, Z. Bingol, C. Firtina, L. Subramanian, J. S. Kim, R. Ausavarungnirun, M. Alser, J. Gomez-Luna, A. Boroumand, A. Nori, A. Scibisz, S. Subramoney, C. Alkan, S. Ghose, and O. Mutlu, "GenASM: A high-performance, low-power approximate string matching acceleration framework for genome sequence analysis," in *Proc. 53rd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2020, pp. 951–966.
- [61] A. Boroumand, "Practical mechanisms for reducing processor-memory data movement in modern workloads," Ph.D. dissertation, Dept. Elect. Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2020.
- [62] D. Lee, S. Ghose, G. Pekhimenko, S. Khan, and O. Mutlu, "Simultaneous multi-layer access: Improving 3D-stacked memory bandwidth at low cost," *ACM Trans. Archit. Code Optim.*, vol. 12, no. 4, pp. 1–29, 2016.
- [63] A. Boroumand, S. Ghose, B. Lucia, K. Hsieh, K. Malladi, H. Zheng, and O. Mutlu, "LazyPIM: Efficient cache coherence mechanism for processing-in-memory," in *Proc. CAL*, 2017, pp. 46–50.
- [64] D. Zhang, N. Jayasena, A. Lysashevsky, J. L. Greathouse, L. Xu, and M. Ignatowski, "TOP-PIM: Throughput-oriented programmable processing in memory," in *Proc. 23rd Int. Symp. High-Perform. Parallel Distrib. Comput. (HPDC)*, 2014, pp. 85–98.
- [65] M. Gao and C. Kozyrakis, "HRL: Efficient and flexible reconfigurable logic for near-data processing," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Mar. 2016, pp. 126–137.
- [66] M. Drumond, A. Daglis, N. Mirzadeh, D. Ustiugov, J. Picorel, B. Falsafi, B. Grot, and D. Pnevmatikatos, "The Mondrian data engine," in *Proc. 44th Annu. Int. Symp. Comput. Archit.*, Jun. 2017, pp. 639–651.
- [67] P. C. Santos, G. F. Oliveira, D. G. Tome, M. A. Z. Alves, E. C. Almeida, and L. Carro, "Operand size reconfiguration for big data processing in memory," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2017, pp. 710–715.
- [68] G. F. Oliveira, P. C. Santos, M. A. Alves, and L. Carro, "NIM: An HMC-based machine for neuron computation," in *Proc. ARC*, 2017, pp. 28–35.
- [69] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "TETRIS: Scalable and efficient neural network acceleration with 3D memory," in *Proc. ASPLOS*, 2017, pp. 751–764.
- [70] D. Kim, J. Kung, S. Chai, S. Yalamanchili, and S. Mukhopadhyay, "Neurocube: A programmable digital neuromorphic architecture with high-density 3D memory," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 380–392.
- [71] P. Gu, S. Li, D. Stow, R. Barnes, L. Liu, Y. Xie, and E. Kursun, "Leveraging 3D technologies for hardware security: Opportunities and challenges," in *Proc. 26th Ed. Great Lakes Symp. (VLSI)*, May 2016, pp. 347–352.
- [72] D. U. Lee, K. W. Kim, K. W. Kim, H. Kim, J. Y. Kim, Y. J. Park, J. H. Kim, D. S. Kim, H. B. Park, J. W. Shin, J. H. Cho, K. H. Kwon, M. J. Kim, J. Lee, K. W. Park, B. Chung, and S. Hong, "A 1.2 V 8 Gb 8-channel 128 GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29 nm process and TSV," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 432–433.
- [73] Hybrid Memory Cube Consortium. *Hybrid Memory Cube Specification Rev. 2.0*. Accessed: Jul. 12, 2021. [Online]. Available: <http://www.hybridmemorycube.org/>
- [74] A. Boroumand, S. Ghose, M. Patel, H. Hassan, B. Lucia, R. Ausavarungnirun, K. Hsieh, N. Hajinazar, K. T. Malladi, H. Zheng, and O. Mutlu, "CoNDA: Efficient cache coherence support for near-data accelerators," in *Proc. 46th Int. Symp. Comput. Archit.*, Jun. 2019, pp. 629–642.
- [75] Q. Zhu, T. Graf, H. E. Sumbul, L. Pileggi, and F. Franchetti, "Accelerating sparse matrix-matrix multiplication with 3D-stacked logic-in-memory hardware," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2013, pp. 1–6.
- [76] S. H. Pugsley, J. Jestes, H. Zhang, R. Balasubramonian, V. Srinivasan, A. Buyuktosunoglu, A. Davis, and F. Li, "NDC: Analyzing the impact of 3D-stacked memory+logic devices on MapReduce workloads," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Mar. 2014, pp. 190–200.
- [77] A. Farmahini-Farahani, J. H. Ahn, K. Morrow, and N. S. Kim, "NDA: Near-DRAM acceleration architecture leveraging commodity DRAM devices and standard memory modules," in *Proc. IEEE 21st Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2015, pp. 283–295.
- [78] G. H. Loh, N. Jayasena, M. Oskin, M. Nutter, D. Roberts, M. Meswani, D. P. Zhang, and M. Ignatowski, "A processing in memory taxonomy and a case for studying fixed-function PIM," in *Proc. WoNDP*, 2013, pp. 1–4.
- [79] A. Pattnaik, X. Tang, A. Jog, O. Kayiran, A. K. Mishra, M. T. Kandemir, O. Mutlu, and C. R. Das, "Scheduling techniques for GPU architectures with processing-in-memory capabilities," in *Proc. Int. Conf. Parallel Architectures Compilation*, Sep. 2016, pp. 31–44.
- [80] B. Akin, F. Franchetti, and J. C. Hoe, "Data reorganization in memory using 3D-stacked DRAM," in *Proc. 42nd Annu. Int. Symp. Comput. Archit.*, Jun. 2015, pp. 131–143.
- [81] A. Augusta and S. Idreos, "JAFAR: Near-data processing for databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2015, pp. 2069–2070.
- [82] J. H. Lee, J. Sim, and H. Kim, "BSSync: Processing near memory for machine learning workloads with bounded staleness consistency models," in *Proc. Int. Conf. Parallel Archit. Compilation (PACT)*, Oct. 2015, pp. 241–252.
- [83] F. Devaux, "The true processing in memory accelerator," in *Proc. IEEE Hot Chips 31 Symp. (HCS)*, Aug. 2019, pp. 1–24.
- [84] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 27–39.
- [85] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: A convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 14–26.
- [86] V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, "Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2017, pp. 273–287.
- [87] V. Seshadri and O. Mutlu, "In-DRAM bulk bitwise execution engine," 2019, *arXiv:1905.09822*. [Online]. Available: <http://arxiv.org/abs/1905.09822>
- [88] S. Li, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "DRISA: A DRAM-based reconfigurable *in-situ* accelerator," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2017, pp. 288–301.
- [89] V. Seshadri, M. A. Kozuch, T. C. Mowry, Y. Kim, C. Fallin, D. Lee, R. Ausavarungnirun, G. Pekhimenko, Y. Luo, O. Mutlu, and P. B. Gibbons, "RowClone: Fast and energy-efficient in-DRAM bulk data copy and initialization," in *Proc. 46th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2013, pp. 185–197.
- [90] V. Seshadri and O. Mutlu, "The processing using memory paradigm: In-DRAM bulk copy, initialization, bitwise AND and OR," 2016, *arXiv:1610.09603*. [Online]. Available: <http://arxiv.org/abs/1610.09603>
- [91] Q. Deng, L. Jiang, Y. Zhang, M. Zhang, and J. Yang, "DrAcc: A DRAM based accelerator for accurate CNN inference," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.
- [92] X. Xin, Y. Zhang, and J. Yang, "ELP2IM: Efficient and low power bitwise operation processing in DRAM," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 303–314.
- [93] L. Song, Y. Zhuo, X. Qian, H. Li, and Y. Chen, "GraphR: Accelerating graph processing using ReRAM," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 531–543.
- [94] L. Song, X. Qian, H. Li, and Y. Chen, "PipeLayer: A pipelined ReRAM-based accelerator for deep learning," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 541–552.
- [95] F. Gao, G. Tziantzioulis, and D. Wentzlauff, "ComputeDRAM: In-memory compute using off-the-shelf DRAMs," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 100–113.
- [96] C. Eckert, X. Wang, J. Wang, A. Subramanian, R. Iyer, D. Sylvester, D. Blaauw, and R. Das, "Neural cache: Bit-serial in-cache acceleration of deep neural networks," in *Proc. ACM/IEEE 45th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2018, pp. 383–396.
- [97] S. Aga, S. Jeloka, A. Subramanian, S. Narayanasamy, D. Blaauw, and R. Das, "Compute caches," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 481–492.
- [98] D. Fujiki, S. Mahlke, and R. Das, "Duality cache for data parallel acceleration," in *Proc. 46th Int. Symp. Comput. Archit.*, Jun. 2019, pp. 397–410.

- [99] I. Fernandez, R. Quisilant, E. Gutiérrez, O. Plata, C. Giannoula, M. Alser, J. Gómez-Luna, and O. Mutlu, "NATSA: A near-data processing accelerator for time series analysis," in *Proc. IEEE 38th Int. Conf. Comput. Design (ICCD)*, Oct. 2020, pp. 120–129.
- [100] N. Hajinazar, G. F. Oliveira, S. Gregorio, J. D. Ferreira, N. M. Ghiasi, M. Patel, M. Alser, S. Ghose, J. Gómez-Luna, and O. Mutlu, "SIM-DRAM: A framework for bit-serial SIMD processing using DRAM," in *Proc. 26th ACM Int. Conf. Architectural Program. Lang. Operating Syst.*, Apr. 2021, pp. 329–345.
- [101] C. Giannoula, N. Vijaykumar, N. Papadopoulou, V. Karakostas, I. Fernandez, J. Gómez-Luna, L. Orosa, N. Koziris, G. Goumas, and O. Mutlu, "SynCron: Efficient synchronization support for near-data-processing architectures," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Feb. 2021, pp. 263–276.
- [102] A. Boroumand, S. Ghose, G. F. Oliveira, and O. Mutlu, "Polynesia: Enabling effective hybrid transactional/analytical databases with specialized hardware/software co-design," 2021, *arXiv:2103.00798*. [Online]. Available: <http://arxiv.org/abs/2103.00798>
- [103] A. Boroumand, S. Ghose, B. Akin, R. Narayanaswami, G. F. Oliveira, X. Ma, E. Shiu, and O. Mutlu, "Mitigating edge machine learning inference bottlenecks: An empirical study on accelerating Google edge models," 2021, *arXiv:2103.00768*. [Online]. Available: <http://arxiv.org/abs/2103.00768>
- [104] G. F. Oliveira, P. C. Santos, M. A. Z. Alves, and L. Carro, "A generic processing in memory cycle accurate simulator under hybrid memory cube architecture," in *Proc. Int. Conf. Embedded Comput. Syst., Archit., Modeling, Simulation (SAMOS)*, Jul. 2017, pp. 54–61.
- [105] J. S. Kim, M. Patel, H. Hassan, L. Orosa, and O. Mutlu, "D-RaNGe: Using commodity DRAM devices to generate true random numbers with low latency and high throughput," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2019, pp. 582–595.
- [106] J. S. Kim, M. Patel, H. Hassan, and O. Mutlu, "The DRAM latency PUF: Quickly evaluating physical unclonable functions by exploiting the latency-reliability tradeoff in modern commodity DRAM devices," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 194–207.
- [107] M. Besta, R. Kanakagiri, G. Kwasniewski, R. Ausavarungrun, J. Beránek, K. Kanellopoulos, K. Janda, Z. Vonarburg-Shmaria, L. Gianinazzi, I. Stefan, J. G. Luna, M. Copik, L. Kapp-Schwoerer, S. D. Girolamo, M. Konieczny, O. Mutlu, and T. Hoefler, "SISA: Set-centric instruction set architecture for graph mining on processing-in-memory systems," 2021, *arXiv:2104.07582*. [Online]. Available: <http://arxiv.org/abs/2104.07582>
- [108] J. Dinis Ferreira, G. Falcao, J. Gómez-Luna, M. Alser, L. Orosa, M. Sadrosadati, J. S. Kim, G. F. Oliveira, T. Shahroodi, A. Nori, and O. Mutlu, "PLUTO: In-DRAM lookup tables to enable massively parallel general-purpose computation," 2021, *arXiv:2104.07699*. [Online]. Available: <http://arxiv.org/abs/2104.07699>
- [109] V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, "Buddy-RAM: Improving the performance and efficiency of bulk bitwise operations using DRAM," 2016, *arXiv:1611.09988*. [Online]. Available: <http://arxiv.org/abs/1611.09988>
- [110] A. Boroumand, S. Ghose, M. Patel, H. Hassan, B. Lucia, N. Hajinazar, K. Hsieh, K. T. Malladi, H. Zheng, and O. Mutlu, "LazyPIM: Efficient support for cache coherence in Processing-in-Memory architectures," 2017, *arXiv:1706.03162*. [Online]. Available: <http://arxiv.org/abs/1706.03162>
- [111] J. S. Kim, D. Senol, H. Xin, D. Lee, S. Ghose, M. Alser, H. Hassan, O. Ergin, C. Alkan, and O. Mutlu, "GRIM-filter: Fast seed filtering in read mapping using emerging memory technologies," 2017, *arXiv:1708.04329*. [Online]. Available: <http://arxiv.org/abs/1708.04329>
- [112] S. Ghose, K. Hsieh, A. Boroumand, R. Ausavarungrun, and O. Mutlu, "Enabling the adoption of Processing-in-memory: Challenges, mechanisms, future research directions," 2018, *arXiv:1802.00320*. [Online]. Available: <http://arxiv.org/abs/1802.00320>
- [113] V. Seshadri, Y. Kim, C. Fallin, D. Lee, R. Ausavarungrun, G. Pekhimenko, Y. Luo, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "RowClone: Accelerating data movement and initialization using DRAM," 2018, *arXiv:1805.03502*. [Online]. Available: <http://arxiv.org/abs/1805.03502>
- [114] S. Ghose, A. Boroumand, J. S. Kim, J. Gómez-Luna, and O. Mutlu, "A workload and programming ease driven perspective of processing-in-memory," 2019, *arXiv:1907.12947*. [Online]. Available: <http://arxiv.org/abs/1907.12947>
- [115] A. Olgun, M. Patel, A. G. Yağlıkcı, H. Luo, J. S. Kim, F. N. Bostancı, N. Vijaykumar, O. Ergin, and O. Mutlu, "QUAC-TRNG: High-throughput true random number generation using quadruple row activation in commodity DRAM chips," in *Proc. ACM/IEEE 48th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 1–15.
- [116] M. Gokhale, B. Holmes, and K. Iobst, "Processing in memory: The terasys massively parallel PIM array," *Computer*, vol. 28, no. 4, pp. 23–31, 1995.
- [117] R. Balasubramonian, J. Chang, T. Manning, J. H. Moreno, R. Murphy, R. Nair, and S. Swanson, "Near-data processing: Insights from a MICRO-46 workshop," *IEEE Micro*, vol. 34, no. 4, pp. 36–42, Jul. 2014.
- [118] R. Nair, "Evolution of memory architecture," *Proc. IEEE*, vol. 103, no. 8, pp. 1331–1345, Jul. 2015.
- [119] E. Azarkhish, D. Rossi, I. Loi, and L. Benini, "A case for near memory computation inside the smart memory cube," in *Proc. EMS*, 2016, pp. 1–2.
- [120] J. D. McCalpin, "Memory bandwidth and machine balance in current high performance computers," *IEEE Comput. Soc. Tech. Committee Comput. Archit. (TCCA) Newslett.*, vol. 2, pp. 19–25, Dec. 1995.
- [121] D. Lee, L. Subramanian, R. Ausavarungrun, J. Choi, and O. Mutlu, "Decoupled direct memory access: Isolating CPU and IO traffic by leveraging a dual-data-port DRAM," in *Proc. Int. Conf. Parallel Archit. Compilation (PACT)*, Oct. 2015, pp. 174–187.
- [122] M. Zhang, Y. Zhuo, C. Wang, M. Gao, Y. Wu, K. Chen, C. Kozyrakis, and X. Qian, "GraphP: Reducing communication for PIM-based graph processing with efficient data partition," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 544–557.
- [123] S. Angizi, J. Sun, W. Zhang, and D. Fan, "GraphS: A graph processing accelerator leveraging SOT-MRAM," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 378–383.
- [124] K. K. Matam, G. Koo, H. Zha, H.-W. Tseng, and M. Annavam, "GraphSSD: Graph semantics aware SSD," in *Proc. ISCA*, 2019, pp. 116–128.
- [125] S. Angizi and D. Fan, "GraphiDe: A graph processing accelerator leveraging in-DRAM-computing," in *Proc. Great Lakes Symp. (VLSI)*, May 2019, pp. 45–50.
- [126] Y. Zhuo, C. Wang, M. Zhang, R. Wang, D. Niu, Y. Wang, and X. Qian, "GraphQ: Scalable PIM-based graph processing," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 712–725.
- [127] K. K. Chang, P. J. Nair, D. Lee, S. Ghose, M. K. Qureshi, and O. Mutlu, "Low-cost inter-linked subarrays (LISA): Enabling fast inter-subarray data movement in DRAM," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Mar. 2016, pp. 568–580.
- [128] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, and Y. Xie, "Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories," in *Proc. 53rd Annu. Design Autom. Conf.*, Jun. 2016, pp. 1–6.
- [129] S. H. S. Rezaei, M. Modarressi, R. Ausavarungrun, M. Sadrosadati, O. Mutlu, and M. Daneshmand, "NoM: Network-on-memory for inter-bank data transfer in highly-banked memories," *IEEE Comput. Archit. Lett.*, vol. 19, no. 1, pp. 80–83, Jun. 2020.
- [130] V. Seshadri, K. Hsieh, A. Boroumand, D. Lee, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, "Fast bulk bitwise AND and OR in DRAM," *IEEE Comput. Archit. Lett.*, vol. 14, no. 2, pp. 127–131, May 2015.
- [131] S. Williams, A. Waterman, and D. Patterson, "Roofline: An insightful visual performance model for multicore architectures," *Commun. ACM*, vol. 52, no. 4, pp. 65–76, 2009.
- [132] E. Azarkhish, D. Rossi, I. Loi, and L. Benini, "Neurostream: Scalable and energy efficient deep learning with smart memory cubes," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 2, pp. 420–434, Sep. 2018.
- [133] L. Ke, U. Gupta, B. Y. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H. H. S. Lee, and M. Li, "RecNMP: Accelerating personalized recommendation with near-memory processing," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May 2020, pp. 790–803.
- [134] B. Asgari, S. Mukhopadhyay, and S. Yalamanchili, "MAHASIM: Machine-learning hardware acceleration using a software-defined intelligent memory system," *J. Signal Process. Syst.*, vol. 93, no. 6, pp. 659–675, Feb. 2020.
- [135] N. Kim, J. Ahn, S. Hong, H. Chafi, and K. Choi, "How much computation power do you need for near-data processing in cloud?" in *Proc. ASBD*, 2017.

- [136] S. Liang, Y. Wang, C. Liu, H. Li, and X. Li, "InS-DLA: An in-SSD deep learning accelerator for near-data processing," in *Proc. 29th Int. Conf. Field Program. Log. Appl. (FPL)*, Sep. 2019, pp. 173–179.
- [137] A. O. Glova, I. Akgun, S. Li, X. Hu, and Y. Xie, "Near-data acceleration of privacy-preserving biomarker search with 3D-stacked memory," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 800–805.
- [138] P. Gu, X. Xie, S. Li, D. Niu, H. Zheng, K. T. Malladi, and Y. Xie, "DLUX: A LUT-based near-bank accelerator for data center deep learning training workloads," in *Proc. TCAD*, 2020, pp. 1586–1599.
- [139] G. Singh, D. Diamantopoulos, C. Hagleitner, J. Gomez-Luna, S. Stuijk, O. Mutlu, and H. Corporaal, "NERO: A near high-bandwidth memory stencil accelerator for weather prediction modeling," in *Proc. 30th Int. Conf. Field-Program. Log. Appl. (FPL)*, Aug. 2020, pp. 9–17.
- [140] J. Gómez-Luna, I. El Hajj, I. Fernández, C. Giannoula, G. F. Oliveira, and O. Mutlu, "Benchmarking a new paradigm: An experimental analysis of a real processing-in-memory architecture," 2021, *arXiv:2105.03814*. [Online]. Available: <http://arxiv.org/abs/2105.03814>
- [141] M. Radulovic, D. Zivanovic, D. Ruiz, B. R. de Supinski, S. A. McKee, P. Radojković, and E. Ayguadé, "Another trip to the wall: How much will stacked DRAM benefit HPC?" in *Proc. MEMSYS*, 2015, pp. 31–36.
- [142] L. Yavits, R. Kaplan, and R. Ginosar, "GIRAF: General purpose in-storage resistive associative framework," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 2, pp. 276–287, Feb. 2021.
- [143] J. M. Herruzo, I. Fernandez, S. González-Navarro, and O. Plata, "Enabling fast and energy-efficient FM-index exact matching using processing-near-memory," *J. Supercomput.*, vol. 77, no. 9, pp. 10226–10251, Sep. 2021.
- [144] B. Asgari, R. Hadidi, J. Cao, D. E. Shim, S.-K. Lim, and H. Kim, "FAFNIR: Accelerating sparse gathering by using efficient near-memory intelligent reduction," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Feb. 2021, pp. 908–920.
- [145] H. Kim, H. Kim, S. Yalamanchili, and A. F. Rodrigues, "Understanding energy aspects of processing-near-memory for HPC workloads," in *Proc. Int. Symp. Memory Syst.*, Oct. 2015, pp. 276–282.
- [146] L. Nai and H. Kim, "Instruction offloading with HMC 2.0 standard: A case study for graph traversals," in *Proc. Int. Symp. Memory Syst.*, Oct. 2015, pp. 258–261.
- [147] H. Lim and G. Park, "Triple engine processor (TEP): A heterogeneous near-memory processor for diverse kernel operations," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 4, pp. 1–25, 2017.
- [148] H. Kim, R. Hadidi, L. Nai, H. Kim, N. Jayasena, Y. Eckert, O. Kayiran, and G. Loh, "CODA: Enabling co-location of computation and data for multiple GPU systems," *ACM Trans. Archit. Code Optim.*, vol. 15, no. 3, pp. 1–23, 2018.
- [149] B. Hong, G. Kim, J. H. Ahn, Y. Kwon, H. Kim, and J. Kim, "Accelerating linked-list traversal through near-data processing," in *Proc. Int. Conf. Parallel Archit. Compilation*, Sep. 2016, pp. 113–124.
- [150] M. Gries, P. Cabré, and J. Gago, "Performance evaluation and feasibility study of near-data processing on DRAM modules (DIMM-NDP) for scientific applications," Huawei Technol., Munich Res. Center, Germany, Tech. Rep. MRC-2019-04-15-R1, Apr. 2019.
- [151] M. Hashemi, E. Ebrahimi, O. Mutlu, and Y. N. Patt, "Accelerating dependent cache misses with an enhanced memory controller," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 444–455.
- [152] C. Chou, P. Nair, and M. K. Qureshi, "Reducing refresh power in mobile devices with morphable ECC," in *Proc. 45th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, Jun. 2015, pp. 355–366.
- [153] Y. Kim, D. Han, O. Mutlu, and M. Harchol-Balter, "ATLAS: A scalable and high-performance scheduling algorithm for multiple memory controllers," in *Proc. 16th Int. Symp. High-Perform. Comput. Archit.*, Jan. 2010, pp. 1–12.
- [154] Y. Kim, M. Papamichael, O. Mutlu, and M. Harchol-Balter, "Thread cluster memory scheduling: Exploiting differences in memory access behavior," in *Proc. 43rd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2010, pp. 65–76.
- [155] S. P. Muralidhara, L. Subramanian, O. Mutlu, M. Kandemir, and T. Moscibroda, "Reducing memory interference in multicore systems via application-aware memory channel partitioning," in *Proc. MICRO*, 2011, pp. 374–385.
- [156] L. Subramanian, D. Lee, V. Seshadri, H. Rastogi, and O. Mutlu, "BLISS: Balancing performance, fairness and complexity in memory access scheduling," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 10, pp. 3071–3087, Oct. 2016.
- [157] H. Usui, L. Subramanian, K. K.-W. Chang, and O. Mutlu, "DASH: Deadline-aware high-performance memory scheduler for heterogeneous systems with hardware accelerators," *ACM Trans. Archit. Code Optim.*, vol. 12, no. 4, pp. 1–28, 2016.
- [158] SAFARI Research Group. *DAMOV Benchmark Suite and Simulation Framework*. Accessed: Jul. 12, 2021. [Online]. Available: <https://github.com/CMU-SAFARI/DAMOV>
- [159] D. Sanchez and C. Kozyrakis, "ZSim: Fast and accurate microarchitectural simulation of thousand-core systems," in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 475–486.
- [160] Y. Kim, W. Yang, and O. Mutlu, "Ramulator: A fast and extensible DRAM simulator," *IEEE Comput. Archit. Lett.*, vol. 15, no. 1, pp. 45–49, Mar. 2016.
- [161] Intel Corporation. (2021). *Intel VTune Profiler User Guide*. [Online]. Available: <https://software.intel.com/content/www/us/en/develop/documentation/vtune-help/>
- [162] A. Yasin, "A top-down method for performance analysis and counters architecture," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Mar. 2014, pp. 35–44.
- [163] Intel Corporation. *Intel VTune Profiler User Guide*. Accessed: Jul. 12, 2021. [Online]. Available: <https://software.intel.com/content/www/us/en/develop/documentation/vtune-help/top/reference/cpu-metrics-reference/memory-bound.html>
- [164] T. M. Conte and W.-M. W. Hwu, "Benchmark characterization," *IEEE Comput.*, vol. 24, no. 1, pp. 48–56, Jan. 1991.
- [165] M. S. Johnstone and P. R. Wilson, "The memory fragmentation problem: Solved?" in *Proc. ISMM*, 1998, pp. 26–36.
- [166] J. Weinberg, M. O. McCracken, E. Strohmaier, and A. Snavely, "Quantifying locality in the memory access patterns of HPC applications," in *Proc. ACM/IEEE SC Conf. (SC05)*, 2005, p. 50.
- [167] Y. S. Shao and D. Brooks, "ISA-independent workload characterization and its implications for specialized architectures," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Apr. 2013, pp. 245–255.
- [168] Y. Zhong, X. Shen, and C. Ding, "Program locality analysis using reuse distance," *ACM Trans. Program. Lang. Syst.*, vol. 31, no. 6, pp. 1–39, 2009.
- [169] X. Gu, I. Christopher, T. Bai, C. Zhang, and C. Ding, "A component model of spatial locality," in *Proc. Int. Symp. Memory Manage. (ISMM)*, 2009, pp. 99–108.
- [170] J. C. Beard and J. Randall, "Eliminating dark bandwidth: A data-centric view of scalable, efficient performance, post-Moore," in *Proc. HiPC*, 2017, pp. 106–114.
- [171] S. Lloyd and M. Gokhale, "In-memory data rearrangement for irregular, data-intensive computing," *IEEE Comput.*, vol. 48, no. 8, pp. 18–25, Aug. 2015.
- [172] T. M. Conte and W.-M. W. Hwu, "A brief survey of benchmark usage in the architecture community," *ACM SIGARCH Comput. Archit. News*, vol. 19, no. 4, pp. 37–44, 1991.
- [173] T. M. Conte and W.-M. W. Hwu, "Advances in benchmarking techniques: New standards and quantitative metrics," in *Proc. Adv. Comput.*, 1995, pp. 231–253.
- [174] D. Doerfler, J. Deslippe, S. Williams, L. Olikier, B. Cook, T. Kurth, M. Lobet, T. Malas, J.-L. Vay, and H. Vincenti, "Applying the roofline performance model to the Intel xeon phi knights landing processor," in *Proc. HiPC*, 2016, pp. 339–353.
- [175] O. Mutlu and T. Moscibroda, "Stall-time fair memory access scheduling for chip multiprocessors," in *Proc. 40th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2007, pp. 146–160.
- [176] O. Mutlu and T. Moscibroda, "Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems," in *Proc. Int. Symp. Comput. Archit.*, Jun. 2008, pp. 63–74.
- [177] E. Ebrahimi, C. J. Lee, O. Mutlu, and Y. N. Patt, "Fairness via source throttling: A configurable and high-performance fairness substrate for multi-core memory systems," *ACM Sigplan Notices*, vol. 45, no. 3, pp. 335–346, 2010.
- [178] S. Palacharla and R. E. Kessler, "Evaluating stream buffers as a secondary cache replacement," in *Proc. 21st Int. Symp. Comput. Archit.*, 1994, pp. 24–33.
- [179] G. Singh, J. Gómez-Luna, G. Mariani, G. F. Oliveira, S. Corda, S. Stuijk, O. Mutlu, and H. Corporaal, "NAPEL: Near-memory computing application performance prediction via ensemble learning," in *Proc. 56th Annu. Design Autom. Conf.*, Jun. 2019, pp. 1–6.
- [180] P. C. Santos, M. A. Z. Alves, M. Diener, L. Carro, and P. O. A. Navaux, "Exploring cache size and core count tradeoffs in systems with reduced memory access latency," in *Proc. 24th Euromicro Int. Conf. Parallel, Distrib., Netw.-Based Process. (PDP)*, Feb. 2016, pp. 388–392.

- [181] A. Glew, "MLP Yes! ILP No!" *ASPLOS Wild Crazy Idea Session*, vol. 98, Oct. 1998.
- [182] M. K. Qureshi, D. N. Lynch, O. Mutlu, and Y. N. Patt, "A case for MLP-aware cache replacement," in *Proc. 33rd Int. Symp. Comput. Archit. (ISCA)*, 2006, pp. 167–178.
- [183] O. Mutlu, H. Kim, and Y. N. Patt, "Efficient runahead execution: Power-efficient memory latency tolerance," *IEEE Micro*, vol. 26, no. 1, pp. 10–20, Jan. 2006.
- [184] O. Mutlu, H. Kim, and Y. N. Patt, "Techniques for efficient processing in runahead execution engines," in *Proc. 32nd Int. Symp. Comput. Archit. (ISCA)*, 2005, pp. 370–381.
- [185] T.-Y. Yeh and Y. N. Patt, "Two-level adaptive training branch prediction," in *Proc. MICRO*, 1991, pp. 51–61.
- [186] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0," in *Proc. 40th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2007, pp. 3–14.
- [187] R. Ausavarungrinun, C. Fallin, X. Yu, K. K.-W. Chang, G. Nazario, R. Das, G. H. Loh, and O. Mutlu, "Design and evaluation of hierarchical rings with deflection routing," in *Proc. IEEE 26th Int. Symp. Comput. Archit. High Perform. Comput.*, Oct. 2014, pp. 230–237.
- [188] M. S. Papamarcos and J. H. Patel, "A low-overhead coherence solution for multiprocessors with private cache memories," in *Proc. ISCA*, 1984, pp. 348–354.
- [189] S. Srinath, O. Mutlu, H. Kim, and Y. N. Patt, "Feedback directed prefetching: Improving the performance and bandwidth-efficiency of hardware prefetchers," in *Proc. IEEE 13th Int. Symp. High Perform. Comput. Archit.*, Feb. 2007, pp. 63–74.
- [190] M. Gao, G. Ayers, and C. Kozyrakis, "Practical near-data processing for in-memory analytics frameworks," in *Proc. Int. Conf. Parallel Archit. Compilation (PACT)*, Oct. 2015, pp. 113–124.
- [191] A. Sethia, G. Dasika, M. Samadi, and S. Mahlke, "Memory-centric system interconnect design with hybrid memory cubes," in *Proc. 22nd Int. Conf. Parallel Archit. Compilation Techn.*, Oct. 2013, pp. 145–155.
- [192] R. Nair, S. F. Antao, C. Bertolli, P. Bose, J. R. Brunheroto, T. Chen, C.-Y. Cher, C. H. Costa, J. Doi, C. Evangelinos, and B. M. Fleischer, "Active memory cube: A processing-in-memory architecture for exascale systems," *IBM J. Res. Develop.*, vol. 59, pp. 1–17, Apr. 2015.
- [193] E. Lockerman, A. Feldmann, M. Bakhshalipour, A. Stanescu, S. Gupta, D. Sanchez, and N. Beckmann, "Livia: Data-centric computing throughout the memory hierarchy," in *Proc. 25th Int. Conf. Architectural Program. Lang. Operating Syst.*, Mar. 2020, pp. 417–433.
- [194] J. P. C. de Lima, P. C. Santos, M. A. Z. Alves, A. C. S. Beck, and L. Carro, "Design space exploration for PIM architectures in 3D-stacked memories," in *Proc. 15th ACM Int. Conf. Comput. Frontiers*, May 2018, pp. 113–120.
- [195] U. Sirin, A. Yasin, and A. Ailamaki, "A methodology for OLTP micro-architectural analysis," in *Proc. 13th Int. Workshop Data Manage. New Hardw.*, May 2017, pp. 1–10.
- [196] R. Appuswamy, J. Fellay, and N. Chaturvedi, "Sequence alignment through the looking glass," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, May 2018, pp. 257–266.
- [197] Intel Corporation. (2011). *Intel Xeon Processor E3-1240*. [Online]. Available: <http://ark.intel.com/products/52273/>
- [198] Intel Corporation. (2018). *Understanding How General Exploration Works in Intel VTune Amplifier*. [Online]. Available: <https://software.intel.com/en-us/articles/understanding-how-general-exploration-works-in-intel-vtune-amplifier-xe>
- [199] J. Gomez-Luna, I. E. Hajj, L.-W. Chang, V. Garcia-Flores, S. G. de Gonzalo, T. B. Jablin, A. J. Pena, and W.-M. Hwu, "Chai: Collaborative heterogeneous applications for integrated-architectures," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Apr. 2017, pp. 43–54.
- [200] United States Department of Energy. (2014). *CORAL Benchmark Codes*. [Online]. Available: <https://asc.llnl.gov/CORAL-benchmarks/>
- [201] J. A. Stratton, C. Rodrigues, I.-J. Sung, N. Obeid, L.-W. Chang, N. Anssari, G. D. Liu, and W.-M. W. Hwu, "Parboil: A revised benchmark suite for scientific and commercial throughput computing," Univ. Illinois Urbana-Champaign, Champaign, IL, USA, IMPACT Res. Group, Tech. Rep. IMPACT-12-01, 2012.
- [202] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: Characterization and architectural implications," in *Proc. 17th Int. Conf. Parallel Archit. Compilation Techn. (PACT)*, 2008, pp. 72–81.
- [203] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, S.-H. Lee, and K. Skadron, "Rodinia: A benchmark suite for heterogeneous computing," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2009, pp. 44–54.
- [204] S. K. Venkata, I. Ahn, D. Jeon, A. Gupta, C. Louie, S. Garcia, S. Belongie, and M. B. Taylor, "SD-VBS: The San Diego vision benchmark suite," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2009, pp. 55–64.
- [205] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The SPLASH-2 programs: Characterization and methodological considerations," in *Proc. 22nd Annu. Int. Symp. Comput. Archit.*, 1995, pp. 24–36.
- [206] P. R. Luszczek, D. H. Bailey, J. J. Dongarra, J. Kepner, R. F. Lucas, R. Rabenseifner, and D. Takahashi, "The HPC challenge (HPCC) benchmark suite," in *Proc. SC*, 2006, pp. 1188455–1188677.
- [207] J. Dongarra, M. A. Heroux, and P. Luszczek, "HPCG benchmark: A new metric for ranking high performance computing systems," Dept. Elect. Eng. Comput. Sci., Univ. Tennessee, Knoxville, TN, USA, Tech. Rep. UT-EECS-15-736, 2015.
- [208] N. Ahmed, K. Bertels, and Z. Al-Ars, "A comparison of seed-and-extend techniques in modern DNA read alignment algorithms," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 1421–1428.
- [209] C. Balkesen, J. Teubner, G. Alonso, and M. T. Oszu, "Main-memory hash joins on modern processor architectures," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 7, pp. 1754–1766, Jul. 2015.
- [210] J. Gomez Luna, L.-W. Chang, I.-J. Sung, W.-M. Hwu, and N. Guil, "In-place data sliding algorithms for many-core architectures," in *Proc. 44th Int. Conf. Parallel Process.*, Sep. 2015, pp. 210–219.
- [211] N. Sundaram, N. Satish, M. M. A. Patwary, S. R. Dulloor, M. J. Anderson, S. G. Vadlamudi, D. Das, and P. Dubey, "GraphMat: High performance graph analytics made productive," *Proc. VLDB Endowment*, vol. 8, no. 11, pp. 1214–1225, Jul. 2015.
- [212] J. Shun and G. E. Blelloch, "Ligra: A lightweight graph processing framework for shared memory," in *Proc. PPoPP*, 2013, pp. 135–146.
- [213] R. M. Yoo, A. Romano, and C. Kozyrakis, "Phoenix rebirth: Scalable MapReduce on a large-scale shared-memory system," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2009.
- [214] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
- [215] J. Redmon. *DarkNet: Open Source Neural Networks in C*. [Online]. Available: <http://pjreddie.com/darknet>
- [216] R. Hadidi, B. Asgari, B. A. Mudassar, S. Mukhopadhyay, S. Yalamanchili, and H. Kim, "Demystifying the characteristics of 3D-stacked memories: A case study for hybrid memory cube," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2017, pp. 66–75.
- [217] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-MAT: A recursive model for graph mining," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2004, pp. 442–446.
- [218] Center for Discrete Mathematics & Theoretical Computer Science. (2014). *9th DIMACS Implementation Challenge*. [Online]. Available: <http://www.dis.uniroma1.it/challenge9/>
- [219] M. Alser, H. Hassan, H. Xin, O. Ergin, O. Mutlu, and C. Alkan, "GateKeeper: A new hardware architecture for accelerating pre-alignment in DNA short read mapping," *Bioinformatics*, vol. 33, no. 21, pp. 3355–3363, 2017.
- [220] M. Alser, Z. Bingöl, D. S. Cali, J. Kim, S. Ghose, C. Alkan, and O. Mutlu, "Accelerating genome analysis: A primer on an ongoing journey," *IEEE Micro*, vol. 40, no. 5, pp. 65–75, Sep. 2020.
- [221] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, p. 100, 1979.
- [222] E. Azarkhish, "Memory hierarchy design for next generation scalable many-core platforms," Ph.D. dissertation, Dept. Elect., Electron., Inf. Eng., Università di Bologna, Bologna, Italy, 2016.
- [223] S. H. Pugsley, J. Jests, R. Balasubramonian, V. Srinivasan, A. Buyuktosunoglu, A. Davis, and F. Li, "Comparing implementations of near-data computing with in-memory MapReduce workloads," *IEEE Micro*, vol. 34, no. 4, pp. 44–52, Jul. 2014.
- [224] Y. Chou, B. Fahs, and S. Abraham, "Microarchitecture optimizations for exploiting memory-level parallelism," in *Proc. 31st Annu. Int. Symp. Comput. Archit.*, 2004, pp. 76–87.
- [225] J. Tuck, L. Ceze, and J. Torrellas, "Scalable cache miss handling for high memory-level parallelism," in *Proc. 39th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2006, pp. 409–422.
- [226] S. Phadke and S. Narayanasamy, "MLP aware heterogeneous memory system," in *Proc. Design, Autom. Test Eur.*, Mar. 2011, pp. 1–6.
- [227] K. Van Craeynest, S. Eyerhan, and L. Eeckhout, "MLP-aware runahead threads in a simultaneous multithreading processor," in *Proc. HIPEAC*, 2009, pp. 110–124.

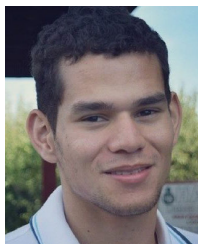
- [228] S. Everman and L. Eeckhout, "A memory-level parallelism aware fetch policy for SMT processors," in *Proc. IEEE 13th Int. Symp. High Perform. Comput. Archit.*, Feb. 2007, pp. 240–249.
- [229] G. Patsilaras, N. K. Choudhary, and J. Tuck, "Efficiently exploiting memory level parallelism on asymmetric coupled cores in the dark silicon era," *ACM Trans. Archit. Code Optim.*, vol. 8, no. 4, pp. 1–21, Jan. 2012.
- [230] Y. Eckert, N. Jayasena, and G. H. Loh, "Thermal feasibility of die-stacked processing in memory," in *Proc. WoNDP*, 2014.
- [231] K. Nilakant, V. Dalibard, A. Roy, and E. Yoneki, "PrefEdge: SSD prefetcher for large-scale graph traversal," in *Proc. Int. Conf. Syst. Storage (SYSTOR)*, 2014, pp. 1–12.
- [232] A. M. Kaushik, G. Pekhimenko, and H. Patel, "Gretch: A hardware prefetcher for graph analytics," in *Proc. TACO*, 2021, pp. 1–25.
- [233] S. Ainsworth and T. M. Jones, "Graph prefetching using data structure knowledge," in *Proc. Int. Conf. Supercomput.*, Jun. 2016.
- [234] A. Basak, S. Li, X. Hu, S. M. Oh, X. Xie, L. Zhao, X. Jiang, and Y. Xie, "Analysis and optimization of the memory hierarchy for graph processing workloads," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2019, pp. 373–386.
- [235] M. Yan, X. Hu, S. Li, A. Basak, H. Li, X. Ma, I. Akgun, Y. Feng, P. Gu, L. Deng, X. Ye, Z. Zhang, D. Fan, and Y. Xie, "Alleviating irregularity in graph analytics acceleration: A hardware/software co-design approach," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 615–628.
- [236] S. T. Srinivasan, R. Rajwar, H. Akkary, A. Gandhi, and M. Upton, "Continual flow pipelines," in *Proc. 11th Int. Conf. Architectural Program. Lang. Operating Syst. (ASPLOS)*, 2004, pp. 107–119.
- [237] M. Annaram, J. M. Patel, and E. S. Davidson, "Data prefetching by dependence graph precomputation," in *Proc. 28th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2001, pp. 52–61.
- [238] J. D. Collins, D. M. Tullsen, H. Wang, and J. P. Shen, "Dynamic speculative precomputation," in *Proc. 34th ACM/IEEE Int. Symp. Microarchitecture (MICRO)*, 2001, pp. 306–317.
- [239] J. Dundas and T. Mudge, "Improving data cache performance by pre-executing instructions under a cache miss," in *Proc. 11th Int. Conf. Supercomput.*, 1997, pp. 68–75.
- [240] O. Mutlu, J. Stark, C. Wilkerson, and Y. N. Patt, "Runahead execution: An effective alternative to large instruction Windows," *IEEE Micro*, vol. 23, no. 6, pp. 20–25, Nov. 2003.
- [241] T. Ramirez, A. Pajuelo, O. J. Santana, and M. Valero, "Runahead threads to improve SMT performance," in *Proc. IEEE 14th Int. Symp. High Perform. Comput. Archit.*, Feb. 2008, pp. 149–158.
- [242] T. Ramirez, A. Pajuelo, O. J. Santana, O. Mutlu, and M. Valero, "Efficient runahead threads," in *Proc. 19th Int. Conf. Parallel Archit. Compilation Techn. (PACT)*, 2010, pp. 443–452.
- [243] W. Zhang, D. M. Tullsen, and B. Calder, "Accelerating and adapting precomputation threads for efficient prefetching," in *Proc. IEEE 13th Int. Symp. High Perform. Comput. Archit.*, 2007, pp. 85–95.
- [244] R. Cooksey, S. Jourdan, and D. Grunwald, "A stateless, content-directed data prefetching mechanism," in *Proc. 10th Int. Conf. Architectural Program. Lang., 10th Int. Conf. Syst. (ASPLOS)*, 2002, pp. 279–290.
- [245] A. Roth and G. S. Sohi, "Effective jump-pointer prefetching for linked data structures," in *Proc. 26th Int. Symp. Comput. Archit.*, 1999, pp. 111–121.
- [246] P. C. Santos, G. F. Oliveira, J. P. Lima, M. A. Z. Alves, L. Carro, and A. C. S. Beck, "Processing in 3D memories to speed up operations on complex data structures," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 897–900.
- [247] S. Ghose, H. Lee, and J. F. Martinez, "Improving memory scheduling via processor-side load criticality information," in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 84–95.
- [248] O. Mutlu. (2015). *Lecture Notes for 18-447 Computer Architecture—Lecture 17: Memory Hierarchy and Caches*. [Online]. Available: <https://course.ece.cmu.edu/~ece447/s15/lib/exe/fetch.php?media=onur-447-spring15-lecture17-memoryhierarchyandcaches-afterlecture.pdf>
- [249] T. L. Johnson, D. A. Connors, M. C. Merten, and W.-M. W. Hwu, "Run-time cache bypassing," *IEEE Trans. Comput.*, vol. 48, no. 12, pp. 1338–1354, Dec. 1999.
- [250] A. Sembrant, E. Hagersten, and D. Black-Schaffer, "The direct-to-data (D2D) cache: Navigating the cache hierarchy with a single lookup," *ACM SIGARCH Comput. Archit. News*, vol. 42, no. 3, pp. 133–144, 2014.
- [251] P.-A. Tsai, N. Beckmann, and D. Sanchez, "Jenga: Software-defined cache hierarchies," in *Proc. 44th Annu. Int. Symp. Comput. Archit.*, Jun. 2017, pp. 652–665.
- [252] V. Seshadri, A. Bhowmick, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "The dirty-block index," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 157–168.
- [253] V. Seshadri, S. Yedkar, H. Xin, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "Mitigating prefetcher-caused pollution using informed caching policies for prefetched blocks," *ACM Trans. Archit. Code Optim.*, vol. 11, no. 4, pp. 1–22, 2015.
- [254] T. L. Johnson and W.-M.-W. Hwu, "Run-time adaptive cache hierarchy management via reference analysis," in *Proc. 24th Annu. Int. Symp. Comput. Archit. (ISCA)*, 1997, pp. 315–326.
- [255] G. Tyson, M. Farrens, J. Matthews, and A. R. Pleszkun, "A modified approach to data cache management," in *Proc. 28th Annu. Int. Symp. Microarchitecture*, Nov. 1995, pp. 93–103.
- [256] G. Memik, G. Reinman, and W. H. Mangione-Smith, "Just say no: Benefits of early cache miss determination," in *Proc. 9th Int. Symp. High-Perform. Comput. Archit. (HPCA)*, 2003, pp. 307–316.
- [257] M. Kharbutli and Y. Solihin, "Counter-based cache replacement and bypassing algorithms," *IEEE Trans. Comput.*, vol. 57, no. 4, pp. 433–447, Feb. 2008.
- [258] S. Gupta, H. Gao, and H. Zhou, "Adaptive cache bypassing for inclusive last level caches," in *Proc. IEEE 27th Int. Symp. Parallel Distrib. Process.*, May 2013, pp. 1243–1253.
- [259] L. Li, D. Tong, Z. Xie, J. Lu, and X. Cheng, "Optimal bypass monitor for high performance last-level caches," in *Proc. 21st Int. Conf. Parallel Archit. Compilation Techn. (PACT)*, 2012, pp. 315–324.
- [260] A. Sridharan and A. Seznec, "Discrete cache insertion policies for shared last level cache management on large multicores," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. (IPDPS)*, May 2016, pp. 822–831.
- [261] D. Lee, Y. Kim, V. Seshadri, J. Liu, L. Subramanian, and O. Mutlu, "Tiered-latency DRAM: A low latency and low cost DRAM architecture," in *Proc. IEEE 19th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2013, pp. 615–626.
- [262] H. Hassan, G. Pekhimenko, N. Vijaykumar, V. Seshadri, D. Lee, O. Ergin, and O. Mutlu, "ChargeCache: Reducing DRAM latency by exploiting row access locality," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Mar. 2016, pp. 581–593.
- [263] H. Hassan, M. Patel, J. S. Kim, A. G. Yağlıkçı, N. Vijaykumar, N. M. Ghiasi, S. Ghose, and O. Mutlu, "CROW: A low-cost substrate for improving DRAM performance, energy efficiency, and reliability," in *Proc. 46th Int. Symp. Comput. Archit.*, Jun. 2019, pp. 129–142.
- [264] Y. Wang, A. Tavakkol, L. Orosa, S. Ghose, N. Mansouri Ghiasi, M. Patel, J. S. Kim, H. Hassan, M. Sadrosadati, and O. Mutlu, "Reducing DRAM latency via charge-level-aware look-ahead partial restoration," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2018, pp. 298–311.
- [265] J. Kim, M. Patel, H. Hassan, and O. Mutlu, "Solar-DRAM: Reducing DRAM access latency by exploiting the variation in local bitlines," in *Proc. IEEE 36th Int. Conf. Comput. Design (ICCD)*, Oct. 2018, pp. 282–291.
- [266] A. Das, H. Hassan, and O. Mutlu, "VRL-DRAM: Improving DRAM performance via variable refresh latency," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.
- [267] K. K. Chang, "Understanding and improving the latency of DRAM-based memory systems," Ph.D. dissertation, Dept. Elect. Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2017.
- [268] H. Hassan, N. Vijaykumar, S. Khan, S. Ghose, K. Chang, G. Pekhimenko, D. Lee, O. Ergin, and O. Mutlu, "SoftMC: A flexible and practical open-source infrastructure for enabling experimental DRAM studies," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 241–252.
- [269] D. Lee, "Reducing DRAM latency at low cost by exploiting heterogeneity," Ph.D. dissertation, Dept. Elect. Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2016.
- [270] H. Luo, T. Shahroodi, H. Hassan, M. Patel, A. G. Yaglikci, L. Orosa, J. Park, and O. Mutlu, "CLR-DRAM: A low-cost DRAM architecture enabling dynamic capacity-latency trade-off," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May 2020, pp. 666–679.
- [271] J. Choi, W. Shin, J. Jang, J. Suh, Y. Kwon, Y. Moon, and L.-S. Kim, "Multiple clone row DRAM: A low latency and area optimized DRAM," *ACM SIGARCH Comput. Archit. News*, vol. 43, pp. 223–234, Jun. 2015.
- [272] Y. H. Son, O. Seongil, Y. Ro, J. W. Lee, and J. H. Ahn, "Reducing memory access latency with asymmetric DRAM bank organizations," in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 380–391.

- [273] Micron. *RLDRAM 2 and 3 Specifications*. Accessed: Jul. 12, 2021. [Online]. Available: <https://www.micron.com/products/dram/rldram-memory>
- [274] Y. Sato, T. Suzuki, T. Aikawa, S. Fujioka, W. Fujieda, H. Kobayashi, H. Ikeda, T. Nagasawa, A. Funyu, Y. Fuji, K. Kawasaki, M. Yamazaki, and M. Taguchi, "Fast cycle RAM (FCRAM): a 20-ns random row access, pipe-lined operating DRAM," in *Proc. Symp. VLSI Circuits. Dig. Tech. Papers*, 1998, pp. 22–25.
- [275] L. Orosa, Y. Wang, M. Sadrosadati, J. S. Kim, M. Patel, I. Puddu, H. Luo, K. Razavi, J. Gomez-Luna, H. Hassan, N. Mansouri-Ghiasi, S. Ghose, and O. Mutlu, "CODIC: A low-cost substrate for enabling custom in-DRAM functionalities and optimizations," in *Proc. ACM/IEEE 48th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 484–497.
- [276] O. Seongil, Y. H. Son, N. S. Kim, and J. H. Ahn, "Row-buffer decoupling: A case for low-latency DRAM microarchitecture," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 337–348.
- [277] R. Ausavarungnirun, K. K.-W. Chang, L. Subramanian, G. H. Loh, and O. Mutlu, "Staged memory scheduling: Achieving high performance and scalability in heterogeneous systems," in *Proc. 39th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2012, pp. 416–427.
- [278] L. Subramanian, D. Lee, V. Seshadri, H. Rastogi, and O. Mutlu, "The blacklisting memory scheduler: Achieving high performance and fairness at low cost," in *Proc. IEEE 32nd Int. Conf. Comput. Design (ICCD)*, Oct. 2014, pp. 8–15.
- [279] L. Subramanian, V. Seshadri, A. Ghosh, S. Khan, and O. Mutlu, "The application slowdown model: Quantifying and controlling the impact of inter-application interference at shared caches and main memory," in *Proc. 48th Int. Symp. Microarchitecture*, Dec. 2015, pp. 62–75.
- [280] L. Subramanian, V. Seshadri, Y. Kim, B. Jaiyen, and O. Mutlu, "MISE: Providing performance predictability and improving fairness in shared main memory systems," in *Proc. IEEE 19th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2013, pp. 639–650.
- [281] S. Rixner, "Memory controller optimizations for web servers," in *Proc. 37th Int. Symp. Microarchitecture (MICRO)*, 2004, pp. 355–366.
- [282] S. Rixner, W. J. Dally, U. J. Kapasi, P. Mattson, and J. D. Owens, "Memory access scheduling," *ACM SIGARCH Comput. Archit. News*, vol. 28, no. 2, pp. 128–138, 2000.
- [283] W. K. Zuravleff and T. Robinson, "Controller for a synchronous DRAM that maximizes throughput by allowing memory requests and commands to be issued out of order," U.S. Patent 5 630 096, May 13, 1997.
- [284] T. Moscibroda and O. Mutlu, "Distributed order scheduling and its application to multi-core dram controllers," in *Proc. 27th ACM Symp. Princ. Distrib. Comput. (PODC)*, 2008, pp. 365–374.
- [285] E. Ebrahimi, R. Miftakhutdinov, C. Fallin, C. J. Lee, J. A. Joao, O. Mutlu, and Y. N. Patt, "Parallel application memory scheduling," in *Proc. MICRO*, 2011, pp. 362–373.
- [286] E. Ipek, O. Mutlu, J. F. Martínez, and R. Caruana, "Self-optimizing memory controllers: A reinforcement learning approach," in *Proc. Int. Symp. Comput. Archit.*, Jun. 2008, pp. 39–50.
- [287] I. Hur and C. Lin, "Adaptive history-based memory schedulers," in *Proc. 37th Int. Symp. Microarchitecture (MICRO)*, 2004, pp. 343–354.
- [288] M. Xie, D. Tong, K. Huang, and X. Cheng, "Improving system throughput and fairness simultaneously in shared memory CMP systems via dynamic bank partitioning," in *Proc. IEEE 20th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2014, pp. 344–355.
- [289] G. L. Yuan, A. Bakhoda, and T. M. Aamodt, "Complexity effective memory access scheduling for many-core accelerator architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture (Micro)*, 2009, pp. 34–44.
- [290] H. Wang, R. Singh, M. J. Schulte, and N. S. Kim, "Memory scheduling towards high-throughput cooperative heterogeneous computing," in *Proc. 23rd Int. Conf. Parallel Archit. Compilation*, Aug. 2014, pp. 331–341.
- [291] C. Kim, D. Burger, and S. W. Keckler, "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches," in *Proc. ASPLOS*, 2002, pp. 211–222.
- [292] T. Moscibroda and O. Mutlu, "A case for bufferless routing in on-chip networks," in *Proc. 36th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2009, pp. 196–207.
- [293] R. Das, O. Mutlu, T. Moscibroda, and C. R. Das, "Application-aware prioritization mechanisms for on-chip networks," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2009, pp. 280–291.
- [294] R. Das, O. Mutlu, T. Moscibroda, and C. R. Das, "Aérgia: Exploiting packet latency slack in on-chip networks," in *Proc. 37th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2010, pp. 106–116.
- [295] G. P. Nychis, C. Fallin, T. Moscibroda, O. Mutlu, and S. Seshan, "On-chip networks from a networking perspective: Congestion and scalability in many-core interconnects," in *Proc. SIGCOMM*, 2012, pp. 407–418.
- [296] M. Besta, S. M. Hassan, S. Yalamanchili, R. Ausavarungnirun, O. Mutlu, and T. Hoefler, "Slim NoC: A low-diameter on-chip network topology for high energy efficiency and scalability," *ACM SIGPLAN Notices*, vol. 53, no. 2, pp. 43–55, 2018.
- [297] C. Fallin, C. Craik, and O. Mutlu, "CHIPPER: A low-complexity bufferless deflection router," in *Proc. IEEE 17th Int. Symp. High Perform. Comput. Archit.*, Feb. 2011, pp. 144–155.
- [298] B. Grot, J. Hestness, S. W. Keckler, and O. Mutlu, "Kilo-NOC: A heterogeneous network-on-chip architecture for scalability and service guarantees," in *Proc. 38th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2011, pp. 401–412.
- [299] L. Benini and G. D. Micheli, "Networks on chip: A new paradigm for systems on chip design," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, 2002, pp. 418–419.
- [300] D. Zhang. *ZSim++*. Accessed: Jul. 12, 2021. [Online]. Available: <https://github.com/dzhang50/zsim-plusplus>
- [301] D. Zhang, X. Ma, M. Thomson, and D. Chiou, "Minnow: Lightweight offload engines for workload management and workload-directed prefetching," *ACM SIGPLAN Notices*, vol. 53, no. 2, pp. 593–607, 2018.
- [302] O. Mutlu. (2020). *Lecture Notes for Digital Design and Computer Architecture—Lecture 15a: Out-of-Order Execution*. [Online]. Available: <https://safari.ethz.ch/digitaltechnik/spring2020/lib/exe/fetch.php?media=onur-digitaldesign-2020-lecture15a-out-of-order-execution-beforelecture.pdf>
- [303] R. Hadidi, L. Nai, H. Kim, and H. Kim, "CAIRO: A compiler-assisted technique for enabling instruction-level offloading of processing-in-memory," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 4, pp. 1–25, 2017.
- [304] H. Asghari-Moghaddam, A. Farmahini-Farahani, K. Morrow, J. H. Ahn, and N. S. Kim, "Near-DRAM acceleration with single-ISA heterogeneous processing in standard memory modules," *IEEE Micro*, vol. 36, no. 1, pp. 24–34, Jan. 2016.
- [305] M. A. Z. Alves, M. Diener, P. C. Santos, and L. Carro, "Large vector extensions inside the HMC," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2016, pp. 1249–1254.
- [306] E. Azarkhish, D. Rossi, I. Loi, and L. Benini, "Design and evaluation of a processing-in-memory architecture for the smart memory cube," in *Proc. ARC*, 2016, pp. 19–31.
- [307] H. Asghari-Moghaddam, Y. H. Son, J. H. Ahn, and N. S. Kim, "Chameleon: Versatile and practical near-DRAM acceleration architecture for large memory systems," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–13.
- [308] J. Picorel, D. Jevdjic, and B. Falsafi, "Near-memory address translation," in *Proc. 26th Int. Conf. Parallel Archit. Compilation Techn. (PACT)*, Sep. 2017, pp. 303–317.
- [309] Z. Liu, I. Calciu, M. Herlihy, and O. Mutlu, "Concurrent data structures for near-memory computing," in *Proc. 29th ACM Symp. Parallelism Algorithms Archit.*, Jul. 2017, pp. 235–245.
- [310] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer-Verlag, 2008.
- [311] E. Azarkhish, D. Rossi, I. Loi, and L. Benini, "High performance AXI-4.0 based interconnect for extensible smart memory cubes," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2015, pp. 1317–1322.
- [312] R. Hadidi, B. Asgari, J. Young, B. Ahmad Mudassar, K. Garg, T. Krishna, and H. Kim, "Performance implications of NoCs on 3D-stacked memories: Insights from the hybrid memory cube," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Apr. 2018, pp. 99–108.
- [313] C. Min, J. Mao, H. Li, and Y. Chen, "NeuralHMC: An efficient HMC-based accelerator for deep neural networks," in *Proc. 24th Asia South Pacific Design Autom. Conf.*, Jan. 2019, pp. 394–399.
- [314] G. Dai, T. Huang, Y. Chi, J. Zhao, G. Sun, Y. Liu, Y. Wang, Y. Xie, and H. Yang, "GraphH: A processing-in-memory architecture for large-scale graph processing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 4, pp. 640–653, Mar. 2018.
- [315] Y. S. Shao, B. Reagen, G.-Y. Wei, and D. Brooks, "Aladdin: A pre-RTL, power-performance accelerator simulator enabling large design space exploration of customized architectures," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 97–108.

- [316] J. Jang, J. Heo, Y. Lee, J. Won, S. Kim, S. J. Jung, H. Jang, T. J. Ham, and J. W. Lee, "Charon: Specialized near-memory processing architecture for clearing dead objects in memory," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 726–739.
- [317] C. Xie, S. L. Song, J. Wang, W. Zhang, and X. Fu, "Processing-in-memory enabled graphics processors for 3D rendering," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 637–648.
- [318] M. Lenjani, P. Gonzalez, E. Sadredini, S. Li, Y. Xie, A. Akel, S. Eilert, M. R. Stan, and K. Skadron, "Fulcrum: A simplified control and access mechanism toward flexible and practical *in-situ* accelerators," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 556–569.
- [319] C. D. Kersey, H. Kim, and S. Yalamanchili, "Lightweight SIMT core designs for intelligent 3D stacked DRAM," in *Proc. Int. Symp. Memory Syst.*, Oct. 2017, pp. 49–59.
- [320] J. Huang, R. R. Puli, P. Majumder, S. Kim, R. Boyapati, K. H. Yum, and E. J. Kim, "Active-routing: Compute on the way for near-data processing," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2019, pp. 674–686.
- [321] S. M. Hassan, S. Yalamanchili, and S. Mukhopadhyay, "Near data processing: Impact and optimization of 3D memory system architecture on the uncore," in *Proc. Int. Symp. Memory Syst.*, Oct. 2015, pp. 11–21.
- [322] Q. Guo, N. Alachiotis, B. Akin, F. Sadi, G. Xu, T. M. Low, L. Pileggi, J. C. Hoe, and F. Franchetti, "3D-stacked memory-side acceleration: Accelerator and system design," in *Proc. WoNDP*, 2014.
- [323] P. Liu, A. Hemani, and K. Paul, "3D-stacked many-core architecture for biological sequence analysis problems," in *Proc. Int. Conf. Embedded Comput. Syst., Archit., Modeling, Simulation (SAMOS)*, Jul. 2015, pp. 1420–1460.
- [324] Q. Zhu, B. Akin, H. E. Sumbul, F. Sadi, J. C. Hoe, L. Pileggi, and F. Franchetti, "A 3D-stacked logic-in-memory accelerator for application-specific data intensive computing," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, Oct. 2013, pp. 1–7.
- [325] S. Padmanabha, A. Lukefahr, R. Das, and S. Mahlke, "Mirage cores: The illusion of many out-of-order cores using in-order hardware," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2017, pp. 745–758.
- [326] C. Fallin, C. Wilkerson, and O. Mutlu, "The heterogeneous block architecture," in *Proc. IEEE 32nd Int. Conf. Comput. Design (ICCD)*, Oct. 2014, pp. 386–393.
- [327] C. Villavieja, J. A. Joao, R. Miftakhutdinov, and Y. N. Patt, "Yoga: A hybrid dynamic VLIW/OoO processor," High Perform. Syst. Group, Univ. Texas Austin, Austin, TX, USA, Tech. Rep. TR-HPS-2014-001, 2014, pp. 305–316.
- [328] M. A. Suleman, M. Hashemi, C. Wilkerson, and Y. N. Patt, "MorphCore: An energy-efficient microarchitecture for high performance ILP and high throughput TLP," in *Proc. 45th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2012, pp. 305–316.
- [329] E. Ipek, M. Kirman, N. Kirman, and J. F. Martinez, "Core fusion: Accommodating software diversity in chip multiprocessors," in *Proc. 34th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2007, pp. 186–197.
- [330] M. A. Suleman, O. Mutlu, J. A. Joao, and Y. N. Patt, "Data marshaling for multi-core architectures," in *Proc. 37th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2010, pp. 441–450.
- [331] P. Petrica, A. M. Izraelievitz, D. H. Albonesi, and C. A. Shoemaker, "Flicker: A dynamically adaptive architecture for power limited multi-core systems," in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 13–23.
- [332] C. Kim, S. Sethumadhavan, M. S. Govindan, N. Ranganathan, D. Gulati, D. Burger, and S. W. Keckler, "Composable lightweight processors," in *Proc. 40th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2007, pp. 381–394.
- [333] A. Lukefahr, S. Padmanabha, R. Das, F. M. Sleiman, R. Dreslinski, T. F. Wenisch, and S. Mahlke, "Composite cores: Pushing heterogeneity into a core," in *Proc. 45th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2012, pp. 317–328.
- [334] M. A. Suleman, O. Mutlu, M. K. Qureshi, and Y. N. Patt, "Accelerating critical section execution with asymmetric multi-core architectures," in *Proc. 14th Int. Conf. Architectural Program. Lang. Operating Syst. (ASPLOS)*, 2009, pp. 253–264.
- [335] J. A. Joao, M. A. Suleman, O. Mutlu, and Y. N. Patt, "Bottleneck identification and scheduling in multithreaded applications," in *Proc. 17th Int. Conf. Architectural Program. Lang. Operating Syst. (ASPLOS)*, 2012, pp. 223–234.
- [336] S. Chaudhry, R. Cypher, M. Ekman, M. Karlsson, A. Landin, S. Yip, H. Zeffner, and M. Tremblay, "Simultaneous speculative threading: A novel pipeline architecture implemented in sun's rock processor," in *Proc. 36th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2009, pp. 484–495.
- [337] J. A. Joao, M. A. Suleman, O. Mutlu, and Y. N. Patt, "Utility-based acceleration of multithreaded applications on asymmetric CMPs," in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 154–165.
- [338] D. Tarjan, M. Boyer, and K. Skadron, "Federation: Repurposing scalar cores for out-of-order instruction issue," in *Proc. 45th Annu. Conf. Design Autom. (DAC)*, 2008, pp. 772–775.
- [339] M. Alipour, S. Kaxiras, D. Black-Schaffer, and R. Kumar, "Delay and bypass: Ready and criticality aware instruction scheduling in out-of-order processors," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 424–434.
- [340] R. Kumar, M. Alipour, and D. Black-Schaffer, "Freeway: Maximizing MLP for slice-out-of-order execution," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2019, pp. 558–569.
- [341] M. Alipour, R. Kumar, S. Kaxiras, and D. Black-Schaffer, "FIFOOrder MicroArchitecture: Ready-aware instruction scheduling for OoO processors," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 716–721.
- [342] R. Kumar, K. I. Farkas, N. P. Jouppi, P. Ranganathan, and D. M. Tullsen, "Single-ISA heterogeneous multi-core architectures: The potential for processor power reduction," in *Proc. 22nd Digit. Avionics Syst. Conf.*, 2003, pp. 81–92.
- [343] R. Kumar, D. M. Tullsen, P. Ranganathan, N. P. Jouppi, and K. I. Farkas, "Single-ISA heterogeneous multi-core architectures for multithreaded workload performance," in *Proc. 31st Annu. Int. Symp. Comput. Archit.*, 2004, pp. 64–75.
- [344] H. Ahmed, P. C. Santos, J. P. C. Lima, R. F. Moura, M. A. Z. Alves, A. C. S. Beck, and L. Carro, "A compiler for automatic selection of suitable processing-in-memory instructions," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 564–569.
- [345] S. Baskaran and J. Sampson, "Decentralized offload-based execution on memory-centric compute cores," in *Proc. Int. Symp. Memory Syst.*, Sep. 2020, pp. 61–76.
- [346] J. Li, X. Wang, A. Tumeo, B. Williams, J. D. Leidel, and Y. Chen, "PIMS: A lightweight processing-in-memory accelerator for stencil computations," in *Proc. Int. Symp. Memory Syst.*, Sep. 2019, pp. 41–52.
- [347] G. Ayers, H. Litz, C. Kozyrakis, and P. Ranganathan, "Classifying memory access patterns for prefetching," in *Proc. 25th Int. Conf. Architectural Program. Lang. Operating Syst.*, Mar. 2020, pp. 513–526.
- [348] J. D. Collins, H. Wang, D. M. Tullsen, C. Hughes, Y.-F. Lee, D. Lavery, and J. P. Shen, "Speculative precomputation: Long-range prefetching of delinquent loads," in *Proc. 28th Annu. Int. Symp. Comput. Archit.*, 2001, pp. 14–25.
- [349] J. González and A. González, "Speculative execution via address prediction and data prefetching," in *Proc. 11th Int. Conf. Supercomputing (ICS)*, 1997, pp. 196–203.
- [350] T.-F. Chen and J.-L. Baer, "Effective hardware-based data prefetching for high-performance processors," *IEEE Trans. Comput.*, vol. 44, no. 5, pp. 609–623, May 1995.
- [351] R. Bera, A. V. Nori, O. Mutlu, and S. Subramoney, "DSPatch: Dual spatial pattern prefetcher," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 531–544.
- [352] M. Bakhshalipour, P. Lotfi-Kamran, and H. Sarbazi-Azad, "Domino temporal data prefetcher," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 131–142.
- [353] J. W. C. Fu, J. H. Patel, and B. L. Janssens, "Stride directed prefetching in scalar processors," in *Proc. 25th Annu. Int. Symp. Microarchitecture (MICRO)*, 1992, pp. 102–110.
- [354] Y. Ishii, M. Inaba, and K. Hiraki, "Access map pattern matching for data cache prefetch," in *Proc. 23rd Int. Conf. Supercomputing (ICS)*, 2009, pp. 499–500.
- [355] M. Hashemi, K. Swersky, J. A. Smith, G. Ayers, H. Litz, J. Chang, C. Kozyrakis, and P. Ranganathan, "Learning memory access patterns," in *Proc. ICML*, 2018, pp. 1919–1928.
- [356] L. Orosa, R. Azevedo, and O. Mutlu, "AVPP: Address-first value-next predictor with value prefetching for improving the efficiency of load value prediction," *ACM Trans. Archit. Code Optim.*, vol. 15, no. 4, pp. 1–30, Jan. 2019.
- [357] A. Jog, O. Kayiran, A. K. Mishra, M. T. Kandemir, O. Mutlu, R. Iyer, and C. R. Das, "Orchestrated scheduling and prefetching for GPGPUs," in *Proc. 40th Annu. Int. Symp. Comput. Archit.*, Jun. 2013, pp. 332–343.
- [358] C. J. Lee, O. Mutlu, V. Narasiman, and Y. N. Patt, "Prefetch-aware memory controllers," *IEEE Trans. Comput.*, vol. 60, no. 10, pp. 1406–1430, Oct. 2011.

- [359] T. M. Austin and G. S. Sohi, "Zero-cycle loads: Microarchitecture support for reducing load latency," in *Proc. 28th Annu. Int. Symp. Microarchitecture*, Nov. 1995, pp. 82–92.
- [360] L. Ceze, K. Strauss, J. Tuck, J. Torrellas, and J. Renau, "CAVA: Using checkpoint-assisted value prediction to hide L2 misses," *ACM Trans. Archit. Code Optim.*, vol. 3, no. 2, pp. 182–208, Jun. 2006.
- [361] D. Kadjo, J. Kim, P. Sharma, R. Panda, P. Gratz, and D. Jimenez, "B-fetch: Branch prediction directed prefetching for chip-multiprocessors," in *Proc. 47th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2014, pp. 623–634.
- [362] N. Kirman, M. Kirman, M. Chaudhuri, and J. F. Martinez, "Checkpointed early load retirement," in *Proc. 11th Int. Symp. High-Perform. Comput. Archit.*, 2005, pp. 16–27.
- [363] D. Joseph and D. Grunwald, "Prefetching using Markov predictors," in *Proc. 24th Annu. Int. Symp. Comput. Archit. (ISCA)*, 1997, pp. 252–263.
- [364] E. Ebrahimi, C. J. Lee, O. Mutlu, and Y. N. Patt, "Prefetch-aware shared resource management for multi-core systems," in *Proc. 38th Annu. Int. Symp. Comput. Archit. (ISCA)*, 2011, pp. 141–152.
- [365] E. Ebrahimi, O. Mutlu, C. J. Lee, and Y. N. Patt, "Coordinated control of multiple prefetchers in multi-core systems," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture (Micro)*, 2009, pp. 316–326.
- [366] C. J. Lee, V. Narasiman, O. Mutlu, and Y. N. Patt, "Improving memory bank-level parallelism in the presence of prefetching," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture (Micro)*, 2009, pp. 327–336.
- [367] C. J. Lee, O. Mutlu, V. Narasiman, and Y. N. Patt, "Prefetch-aware DRAM controllers," in *Proc. 41st IEEE/ACM Int. Symp. Microarchitecture*, Nov. 2008, pp. 200–209.
- [368] M. J. Charney and A. P. Reeves, "Generalized correlation-based hardware prefetching," Cornell Univ., Ithaca, NY, USA, Tech. Rep. EE-CEG-95-1, 1995.
- [369] M. Charney, "Correlation-based hardware prefetching," Ph.D. dissertation, Dept. Elect. Comput. Eng., Cornell Univ., Ithaca, NY, USA, 1995.
- [370] O. Mutlu, H. Kim, D. N. Armstrong, and Y. N. Patt, "Using the first-level caches as filters to reduce the pollution caused by speculative memory references," *Int. J. Parallel Program.*, vol. 33, no. 5, pp. 529–559, Oct. 2005.
- [371] A. Yazdanbakhsh, G. Pekhimenko, B. Thwaites, H. Esmailzadeh, O. Mutlu, and T. C. Mowry, "RFVP: Rollback-free value prediction with safe-to-approximate loads," *ACM Trans. Archit. Code Optim.*, vol. 12, no. 4, pp. 1–26, Jan. 2016.
- [372] R. J. Eickemeyer and S. Vassiliadis, "A load-instruction unit for pipelined processors," *IBM J. Res. Develop.*, vol. 37, no. 4, pp. 547–564, Jul. 1993.
- [373] F. A. Endo, A. Perais, and A. Sez nec, "On the interactions between value prediction and compiler optimizations in the context of EOLE," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 2, pp. 1–24, Jul. 2017.
- [374] M. H. Lipasti and J. P. Shen, "Exceeding the dataflow limit via value prediction," in *Proc. MICRO*, 1996, pp. 226–237.
- [375] M. H. Lipasti, C. B. Wilkerson, and J. P. Shen, "Value locality and load value prediction," in *Proc. ASPLOS*, 1996, pp. 138–147.
- [376] B. Calder, G. Reinman, and D. M. Tullsen, "Selective value prediction," in *Proc. 26th Int. Symp. Comput. Archit.*, 1999, pp. 64–74.
- [377] K. Wang and M. Franklin, "Highly accurate data value prediction using hybrid predictors," in *Proc. 30th Annu. Int. Symp. Microarchitecture*, 1997, pp. 281–290.
- [378] F. Gabbay and A. Mendelson, "Speculative execution based on value prediction," Technion-Israel Inst. Technol., Haifa, Israel, Tech. Rep. 1080, 1996.
- [379] M. Burtscher and B. G. Zorn, "Exploring last n value prediction," in *Proc. Int. Conf. Parallel Architectures Compilation Techn.*, 1999, pp. 66–76.
- [380] C.-Y. Fu, M. D. Jennings, S. Y. Larin, and T. M. Conte, "Value speculation scheduling for high performance processors," in *Proc. 8th Int. Conf. Architectural Program. Lang. Operating Syst. (ASPLOS)*, 1998, pp. 262–271.
- [381] B. Goeman, H. Vandierendonck, and K. de Bosschere, "Differential FCM: Increasing value prediction accuracy by improving table usage efficiency," in *Proc. HPCA 7th Int. Symp. High-Perform. Comput. Archit.*, 2001, pp. 207–216.
- [382] T. Nakra, R. Gupta, and M. L. Soffa, "Global context-based value prediction," in *Proc. 5th Int. Symp. High-Perform. Comput. Archit.*, 1999, pp. 4–12.
- [383] T. Sato and I. Arita, "Low-cost value predictors using frequent value locality," in *Proc. HPCA*, 2002, pp. 106–119.
- [384] Y. Sazeides and J. E. Smith, "The predictability of data values," in *Proc. MICRO*, 1997, pp. 248–258.
- [385] N. Tuck and D. M. Tullsen, "Multithreaded value prediction," in *Proc. 11th Int. Symp. High-Perform. Comput. Archit.*, 2005, pp. 5–15.
- [386] D. M. Tullsen and J. S. Seng, "Storageless value prediction using prior register values," in *Proc. 26th Int. Symp. Comput. Archit.*, 1999, pp. 270–279.
- [387] E. S. Tune, D. M. Tullsen, and B. Calder, "Quantifying instruction criticality," in *Proc. Int. Conf. Parallel Archit. Compilation Techn.*, 2002, pp. 104–113.
- [388] G. Pekhimenko, V. Seshadri, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "Base-delta-immediate compression: Practical data compression for on-chip caches," in *Proc. PACT*, 2012, pp. 377–388.
- [389] J. Dusser, T. Piquet, and A. Sez nec, "Zero-content augmented caches," in *Proc. ICS*, 2009, pp. 46–55.
- [390] J. Yang, Y. Zhang, and R. Gupta, "Frequent value compression in data caches," in *Proc. 33rd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, 2000, pp. 258–265.
- [391] A. R. Alameldeen and D. A. Wood, "Adaptive cache compression for high-performance processors," in *Proc. 31st Annu. Int. Symp. Comput. Archit.*, 2004, pp. 212–223.
- [392] Y. Zhang, J. Yang, and R. Gupta, "Frequent value locality and value-centric data cache design," in *Proc. 9th Int. Conf. Architectural Program. Lang. Operating Syst. (ASPLOS)*, 2000, pp. 150–159.
- [393] G. Pekhimenko, E. Bolotin, N. Vijaykumar, O. Mutlu, T. C. Mowry, and S. W. Keckler, "A case for toggle-aware compression for GPU systems," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Mar. 2016, pp. 188–200.
- [394] G. Pekhimenko, E. Bolotin, M. O'Connor, O. Mutlu, T. C. Mowry, and S. W. Keckler, "Toggle-aware compression for GPUs," *IEEE Comput. Archit. Lett.*, vol. 14, no. 2, pp. 164–168, Jul. 2015.
- [395] G. Pekhimenko, E. Bolotin, M. O'Connor, O. Mutlu, T. C. Mowry, and S. W. Keckler, "Energy-efficient data compression for GPU memory systems," in *Proc. ASPLOS*, 2015.
- [396] G. Pekhimenko, T. Huberty, R. Cai, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "Exploiting compressed block size as an indicator of future reuse," in *Proc. IEEE 21st Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2015, pp. 51–63.
- [397] G. Pekhimenko, V. Seshadri, Y. Kim, H. Xin, O. Mutlu, P. B. Gibbons, M. A. Kozuch, and T. C. Mowry, "Linearly compressed pages: A low-complexity, low-latency main memory compression framework," in *Proc. MICRO*, 2013, pp. 172–184.
- [398] X. Chen, L. Yang, R. P. Dick, L. Shang, and H. Lekatsas, "C-pack: A high-performance microprocessor cache compression algorithm," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 8, pp. 1196–1208, Aug. 2010.
- [399] E. G. Hallnor and S. K. Reinhardt, "A unified compressed memory hierarchy," in *Proc. 11th Int. Symp. High-Perform. Comput. Archit.*, 2005, pp. 201–212.
- [400] D. W. Hammerstrom and E. S. Davidson, "Information content of CPU memory referencing behavior," in *Proc. 4th Annu. Symp. Comput. Archit. (ISCA)*, 1977, pp. 184–192.
- [401] M. M. Islam and P. Stenstrom, "Zero-value caches: Cancelling loads that return zero," in *Proc. 18th Int. Conf. Parallel Architectures Compilation Techn.*, Sep. 2009, pp. 237–245.
- [402] A. Arelakis and P. Stenstrom, "SC2: A statistical compression cache scheme," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 145–156.
- [403] M. Ekman and P. Stenstrom, "A robust main-memory compression scheme," in *Proc. 32nd Int. Symp. Comput. Archit. (ISCA)*, 2005, pp. 74–85.
- [404] N. Vijaykumar, G. Pekhimenko, A. Jog, A. Bhowmick, R. Ausavarungnirun, C. Das, M. Kandemir, T. C. Mowry, and O. Mutlu, "A case for core-assisted bottleneck acceleration in GPUs: Enabling flexible data compression with assist warps," in *Proc. 42nd Annu. Int. Symp. Comput. Archit.*, Jun. 2015, pp. 41–53.
- [405] J. Gaur, A. R. Alameldeen, and S. Subramoney, "Base-victim compression: An opportunistic cache compression architecture," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 317–328.
- [406] J. S. Miguel, J. Albericio, A. Moshovos, and N. E. Jerger, "Doppelgänger: A cache for approximate computing," in *Proc. ISCA*, 2015, pp. 50–61.
- [407] G. F. Oliveira, L. R. Gonçalves, M. Brandalero, A. C. S. Beck, and L. Carro, "Employing classification-based algorithms for general-purpose approximate computing," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.
- [408] N. Vijaykumar, A. Jain, D. Majumdar, K. Hsieh, G. Pekhimenko, E. Ebrahimi, N. Hajinazar, P. B. Gibbons, and O. Mutlu, "A case for richer cross-layer abstractions: Bridging the semantic gap with expressive memory," in *Proc. ACM/IEEE 45th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2018, pp. 207–220.

- [409] A. Yazdanbakhsh, B. Thwaites, H. Esmailzadeh, G. Pekhimenko, O. Mutlu, and T. C. Mowry, "Mitigating the memory bottleneck with approximate load value prediction," *IEEE Des. Test. Comput.*, vol. 33, no. 1, pp. 32–42, Feb. 2016.
- [410] P.-A. Tsai, Y. L. Gan, and D. Sanchez, "Rethinking the memory hierarchy for modern languages," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2018, pp. 203–216.
- [411] M. A. Z. Alves, P. C. Santos, M. Diener, and L. Carro, "Opportunities and challenges of performing vector operations inside the DRAM," in *Proc. Int. Symp. Memory Syst.*, Oct. 2015, pp. 22–28.
- [412] Z. Sura, A. Jacob, T. Chen, B. Rosenburg, O. Sallenave, C. Bertolli, S. Antao, J. Brunheroto, Y. Park, K. O'Brien, and R. Nair, "Data access optimization in a processing-in-memory system," in *Proc. 12th ACM Int. Conf. Comput. Frontiers*, May 2015, pp. 1–8.
- [413] D. P. Zhang, N. Jayasena, A. Lyashkevsky, J. Greathouse, M. Meswani, M. Nutter, and M. Ignatowski, "A new perspective on processing-in-memory architecture design," in *Proc. ACM SIGPLAN Workshop Memory Syst. Perform. Correctness (MSPC)*, 2013, pp. 1–3.
- [414] N. S. Kim, D. Chen, J. Xiong, and W.-M.-W. Hwu, "Heterogeneous computing meets near-memory acceleration and high-level synthesis in the post-Moore era," *IEEE Micro*, vol. 37, no. 4, pp. 10–18, 2017.
- [415] P.-A. Tsai and D. Sanchez, "Compress objects, not cache lines: An object-based compressed memory hierarchy," in *Proc. 24th Int. Conf. Architectural Program. Lang. Operating Syst.*, Apr. 2019, pp. 229–242.
- [416] N. Vijaykumar, E. Ebrahimi, K. Hsieh, P. B. Gibbons, and O. Mutlu, "The locality descriptor: A holistic cross-layer abstraction to express data locality in GPUs," in *Proc. ACM/IEEE 45th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2018, pp. 829–842.
- [417] O. Mutlu, "Intelligent architectures for intelligent computing systems," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Feb. 2021, pp. 318–323.
- [418] G. Narancic, P. Judd, D. Wu, I. Atta, M. Elnacouzi, J. Zebchuk, J. Albericio, N. E. Jerger, A. Moshovos, K. Kutulakos, and S. Gadelrab, "Evaluating the memory system behavior of smartphone workloads," in *Proc. Int. Conf. Embedded Comput. Syst., Archit., Modeling, Simulation (SAMOS XIV)*, Jul. 2014, pp. 83–92.
- [419] K. Yan, X. Zhang, and X. Fu, "Characterizing, modeling, and improving the QoE of mobile devices with low battery level," in *Proc. 48th Int. Symp. Microarchitecture*, Dec. 2015, pp. 713–724.
- [420] D. Shingari, A. Arunkumar, and C.-J. Wu, "Characterization and throttling-based mitigation of memory interference for heterogeneous smartphones," in *Proc. IEEE Int. Symp. Workload Characterization*, Oct. 2015, pp. 22–33.
- [421] C. Kozyrakis, A. Kansal, S. Sankar, and K. Vaid, "Server engineering insights for large-scale online services," *IEEE Micro*, vol. 30, no. 4, pp. 8–19, Jul. 2010.
- [422] A. Yasin, Y. Ben-Asher, and A. Mendelson, "Deep-dive analysis of the data analytics workload in CloudSuite," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2014, pp. 202–211.
- [423] U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H.-S. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, and X. Zhang, "The architectural implications of Facebook's DNN-based personalized recommendation," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 488–501.
- [424] Y. Gan and C. Delimitrou, "The architectural implications of cloud microservices," *IEEE Comput. Archit. Lett.*, vol. 17, no. 2, pp. 155–158, May 2018.
- [425] G. Ayers, J. H. Ahn, C. Kozyrakis, and P. Ranganathan, "Memory hierarchy for web search," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 643–656.
- [426] R. C. Murphy, P. M. Kogge, and A. Rodrigues, "The characterization of data intensive memory workloads on distributed PIM systems," in *Intelligent Memory Systems*. Berlin, Germany: Springer-Verlag, 2001.
- [427] A. Limaye and T. Adegbiya, "A workload characterization of the SPEC CPU2017 benchmark suite," in *Proc. ISPASS*, 2018, pp. 149–158.
- [428] S. Corda, G. Singh, A. J. Awan, R. Jordans, and H. Corporaal, "Memory and parallelism analysis using a platform-independent approach," in *Proc. 22nd Int. Workshop Softw. Compil. Embedded Syst.*, May 2019, pp. 23–26.
- [429] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A case for intelligent RAM," *IEEE Micro*, vol. 17, no. 2, pp. 34–44, 1997.
- [430] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.
- [431] J. Beranek. *Hardware Effects*. Accessed: Jul. 12, 2021. [Online]. Available: <https://github.com/Kobzol/hardware-effects>
- [432] L.-N. Pouchet. *PolyBench: The Polyhedral Benchmark Suite*. Accessed: Jul. 12, 2021. [Online]. Available: <https://www.cs.colostate.edu/~pouchet/software/polybench/>
- [433] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.
- [434] H. Mo, "Mesoscale modeling and direct simulation of explosively dispersed granular materials," Ph.D. dissertation, Dept. Mech. Mechatron., Univ. Waterloo, Waterloo, ON, Canada, 2019.
- [435] S. Thomas, C. Gohkale, E. Tanuwidjaja, T. Chong, D. Lau, S. Garcia, and M. B. Taylor, "CortexSuite: A synthetic brain benchmark suite," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2014, pp. 76–79.
- [436] M. J. Chaisson and G. Tesler, "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory," *BMC Bioinf.*, vol. 13, no. 1, Dec. 2012.
- [437] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," 2013, *arXiv:1303.3997*. [Online]. Available: <http://arxiv.org/abs/1303.3997>
- [438] WebM. *VP8/VP9 Codec SDK*. Accessed: Jul. 12, 2021. [Online]. Available: <https://github.com/ittiamvpx/libvpx>
- [439] C. Balkesen, G. Alonso, J. Teubner, and M. T. Özsu, "Multi-core, main-memory joins: Sort vs. hash revisited," *Proc. VLDB Endowment*, vol. 7, no. 1, pp. 85–96, Sep. 2013.
- [440] Standard Performance Evaluation Corporation. *SPEC CPU2006 Benchmarks*. Accessed: Jul. 12, 2021. [Online]. Available: <http://www.spec.org/cpu2006/>
- [441] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck, "Contextual LSTM (CLSTM) models for large scale NLP tasks," 2016, *arXiv:1602.06291*. [Online]. Available: <http://arxiv.org/abs/1602.06291>
- [442] A. Buluç and J. R. Gilbert, "The combinatorial BLAS: Design, implementation, and applications," *Int. J. High Perform. Comput. Appl.*, vol. 25, no. 4, pp. 496–509, 2011.
- [443] M. Karasikov, H. Mustafa, D. Danciu, M. Zimmermann, C. Barber, G. Ratsch, and A. Kahles, "MetaGraph: Indexing and analysing nucleotide archives at petabase-scale," *bioRxiv*, Jan. 2020, Art. no. 322164.
- [444] Intel Corporation. *Accelerate Fast Math With Intel OneAPI Math Kernel Library*. Accessed: Jul. 12, 2021. [Online]. Available: <https://software.intel.com/content/www/us/en/develop/tools/oneapi/components/onemkl.html>
- [445] Standard Performance Evaluation Corporation. *SPEC CPU2017 Benchmarks*. Accessed: Jul. 12, 2021. [Online]. Available: <http://www.spec.org/cpu2017/>
- [446] D. R. Llanos, "TPCC-UVa: An open-source TPC-C implementation for global performance measurement of computer systems," *ACM SIGMOD Rec.*, vol. 35, no. 4, pp. 6–15, Dec. 2006.
- [447] X. Yu, G. Bezerra, A. Pavlo, S. Devadas, and M. Stonebraker, "Staring into the abyss: An evaluation of concurrency control with one thousand cores," in *Proc. VLDB*, 2014, pp. 209–220.
- [448] M. Naumov, D. Mudigere, H.-J. M. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C. J. Wu, A. G. Azzolini, and D. Dzhulgakov, "Deep learning recommendation model for personalization and recommendation systems," 2019, *arXiv:1906.00091*. [Online]. Available: <http://arxiv.org/abs/1906.00091>
- [449] R. Palomar, J. Gómez-Luna, F. A. Cheikh, J. Olivares-Bueno, and O. J. Elle, "High-performance computation of Bézier surfaces on parallel and heterogeneous platforms," *Int. J. Parallel Program.*, vol. 46, no. 6, pp. 1035–1062, 2018.
- [450] H. Kasture and D. Sanchez, "TailBench: A benchmark suite and evaluation methodology for latency-critical applications," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Sep. 2016, pp. 1–10.
- [451] S. Nalli, S. Haria, M. D. Hill, M. M. Swift, H. Volos, and K. Keeton, "An analysis of persistent memory use with WHISPER," in *Proc. 22nd Int. Conf. Architectural Program. Lang. Operating Syst.*, Apr. 2017, pp. 135–148.
- [452] K. Kara, D. Alistarh, G. Alonso, O. Mutlu, and C. Zhang, "FPGA-accelerated dense linear machine learning: A precision-convergence trade-off," in *Proc. IEEE 25th Annu. Int. Symp. Field-Program. Custom Comput. Mach. (FCCM)*, Apr. 2017, pp. 160–167.



GERALDO F. OLIVEIRA received the B.S. degree in computer science from the Federal University of Viçosa, Viçosa, Brazil, in 2015, and the M.S. degree in computer science from the Federal University of Rio Grande do Sul, Porto Alegre, Brazil, in 2017. He is currently pursuing the Ph.D. degree with ETH Zürich, Zürich, Switzerland, under the supervision of Onur Mutlu. His current research interests include system support for processing-in-memory and processing-using-memory architectures, data-centric accelerators for emerging applications, approximate computing, and emerging memory systems for consumer devices. He has several publications on these topics.



JUAN GÓMEZ-LUNA (Member, IEEE) received the B.S. and M.S. degrees in telecommunication engineering from the University of Seville, Spain, in 2001, and the Ph.D. degree in computer science from the University of Córdoba, Spain, in 2012. From 2005 to 2017, he was a Faculty Member of the University of Córdoba. He is currently a Senior Researcher and a Lecturer with SAFARI Research Group, ETH Zürich. He is the lead author of PrIM (<https://github.com/CMU-SAFARI/prim-benchmarks>), the first publicly-available benchmark suite for a real-world processing-in-memory architecture, and Chai (<https://github.com/chai-benchmarks/chai>), a benchmark suite for heterogeneous systems with CPU/GPU/FPGA. His research interests include processing-in-memory, memory systems, heterogeneous computing, and hardware and software acceleration of medical imaging and bioinformatics.



LOIS OROSA (Member, IEEE) received the Ph.D. degree from the University of Santiago de Compostela. He was a Postdoctoral Researcher with the Institute of Computing, University of Campinas. He was a Visiting Scholar with the University of Illinois at Urbana-Champaign and Universidade NOVA de Lisboa. He acquired industrial experience at several companies. He is a Senior Researcher with the SAFARI Research Group, ETH Zurich. His current research interests include computer architecture, hardware security, memory systems, and machine learning accelerators.



SAUGATA GHOSE (Member, IEEE) received the B.S. degree in computer science from Binghamton University, the B.S. degree in computer engineering from the State University of New York, and the M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University. He is currently an Assistant Professor with the Department of Computer Science, University of Illinois at Urbana-Champaign. Prior to joining University of Illinois Urbana-Champaign, he was a Postdoctoral Researcher and later a Systems Scientist with Carnegie Mellon University. His current research interests include data-oriented computer architectures and systems, new interfaces between systems software and architectures, low-power memory and storage systems, and architectures for emerging platforms and domains. He received the Best Paper Award from DFRWS-EU for work on solid-state drive forensics, in 2017, and the 2019 Wimmer Faculty Fellow at CMU. For more information, please visit his website at <https://ghose.web.illinois.edu/>.



NANDITA VIJAYKUMAR received the M.S. and Ph.D. degrees from Carnegie Mellon University, in 2019, where she was advised by Prof. Onur Mutlu and Prof. Phil Gibbons. She is currently an Assistant Professor with the Computer Science Department, University of Toronto, and the Department of Computer and Mathematical Sciences, University of Toronto Scarborough. She is also affiliated with the Vector Institute for Artificial Intelligence. Before joining the University of Toronto, she was a Research Scientist with the Memory Architecture and Accelerator Laboratory, Intel Labs. She worked with the Systems Group, Computer Science Department, ETH Zurich, as a Visiting Student. In the past, she worked for AMD, Intel, Microsoft, and Nvidia. Her research interests include computer architecture, compilers, and systems with a focus on the interaction between programming models, systems, and architectures. Her current interests include the system-level and programming challenges of robotics, and large-scale machine learning. For more information, please visit her website at <http://www.cs.toronto.edu/~nandita/>.



IVAN FERNANDEZ received the B.S. degree in computer engineering and the M.S. degree in mechatronics engineering from the University of Malaga, in 2017 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include processing in memory, near-data processing, stacked memory architectures, high-performance computing, trans-precision computing, and time series analysis.



MOHAMMAD SADROSADATI received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Sharif University of Technology, Tehran, Iran, in 2012, 2014, and 2019, respectively. From April 2017 to April 2018, he spent one year as an Academic Guest at ETH Zurich, hosted by Prof. Onur Mutlu during his Ph.D. program. He is currently a Postdoctoral Researcher with ETH Zurich, working under the supervision of Prof. Onur Mutlu. His research interests include heterogeneous computing, processing-in-memory, memory systems, and interconnection networks. Due to his achievements and impact on improving the energy efficiency of GPUs, he received Khwarizmi Youth Award, one of the most prestigious awards, as the first laureate, in 2020, to honor and embolden him to keep taking even bigger steps in his research career.



ONUR MUTLU (Fellow, IEEE) received the B.S. degree in computer engineering and psychology from the University of Michigan, Ann Arbor, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin. He is currently a Professor of computer science with ETH Zurich. He is also a Faculty Member with Carnegie Mellon University, where he previously held Strecker Early Career Professorship. He started the Computer Architecture

Group, Microsoft Research, from 2006 to 2009, and held various product and research positions at Intel Corporation, Advanced Micro Devices, VMware, and Google. His current research interests include computer architecture, systems, hardware security, and bioinformatics. A variety of techniques he, along with his group and collaborators, has invented over the years have influenced industry and have been employed in commercial microprocessors and memory/storage systems. He is an ACM Fellow, and an Elected Member

of the Academy of Europe (Academia Europaea). He received the IEEE High Performance Computer Architecture Test of Time Award, the IEEE Computer Society Edward J. McCluskey Technical Achievement Award, ACM SIGARCH Maurice Wilkes Award, the Inaugural IEEE Computer Society Young Computer Architect Award, the Inaugural Intel Early Career Faculty Award, U.S. National Science Foundation Career Award, Carnegie Mellon University Ladd Research Award, and faculty partnership awards from various companies, and a healthy number of best paper or “Top Pick” paper recognitions at various computer systems, architecture, and security venues. His computer architecture and digital logic design course lectures and materials are freely available on YouTube (<https://www.youtube.com/OnurMutluLectures>), and his research group makes a wide variety of software and hardware artifacts freely available online (<https://safari.ethz.ch/>). For more information, please see his webpage at <https://people.inf.ethz.ch/omutlu/>.

...