Panthera: Holistic Memory Management for Big Data Processing over Hybrid Memories

Chenxi Wang
University of California, Los Angeles
USA

Huimin Cui
SKL Computer Architecture, ICT, CAS
University of Chinese Academy of Sciences, China

Ting Cao
Microsoft Research
China

John Zigman
University of Sydney, Australia

Haris Volos
Google, USA

Onur Mutlu
ETH Zürich, Switzerland

Fang Lv
SKL Computer Architecture, ICT, CAS, China

Xiaobing Feng
SKL Computer Architecture, ICT, CAS
University of Chinese Academy of Sciences, China

Guoqing Harry Xu
University of California, Los Angeles
USA

Abstract
To process real-world datasets, modern data-parallel systems often require extremely large amounts of memory, which are both costly and energy-inefficient. Emerging non-volatile memory (NVM) technologies offer high capacity compared to DRAM and low energy compared to SSDs. Hence, NVMs have the potential to fundamentally change the dichotomy between DRAM and durable storage in Big Data processing. However, most Big Data applications are written in managed languages and executed on top of a managed runtime that already performs various dimensions of memory management. Supporting hybrid physical memories adds in a new dimension, creating unique challenges in data replacement.

This paper proposes Panthera, a semantics-aware, fully automated memory management technique for Big Data processing over hybrid memories. Panthera analyzes user programs on a Big Data system to infer their coarse-grained access patterns, which are then passed to the Panthera runtime for efficient data placement and migration. For Big Data applications, the coarse-grained data division information is accurate enough to guide the GC for data layout, which hardly incurs overhead in data monitoring and moving. We implemented Panthera in OpenJDK and Apache Spark. Our extensive evaluation demonstrates that Panthera reduces energy by 37 – 52% at only a 1 – 4% time overhead.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PLDI ’19, June 22–26, 2019, Phoenix, AZ, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6712-7/19/06...
https://doi.org/10.1145/3314221.3314650

CCS Concepts • Hardware → Non-volatile memory;
• Software and its engineering → Memory management;

Keywords hybrid memories, Big Data systems, memory management, garbage collection

ACM Reference Format:

1 Introduction

Modern Big Data computing exemplified by systems such as Spark and Hadoop is extremely memory-intensive. Lack of memory can lead to a range of severe functional and performance issues including out-of-memory crashes, significantly degraded efficiency, or even loss of data upon node failures. Relying completely on DRAM to satisfy the memory need of a data center is costly in many different ways — e.g., large-volume DRAM is expensive and energy-inefficient; furthermore, DRAM’s relatively small capacity dictates that a large number of machines is often needed just to provide sufficient memory, resulting in underutilized CPU resources for workloads that cannot be easily parallelized.

Emerging non-volatile memory (NVM), such as phase change memory (PCM) [31, 53, 56], resistive random-access memory (RRAM) [52], Spin-transfer torque memory (STT-MRAM) [29] or 3D XPoint [4], is a promising technology that, compared to traditional DRAM, provides higher capacity and lower energy consumption. Systems with hybrid memories have therefore received much attention [9–11, 13, 16, 23, 30, 31, 34, 37, 39, 40, 45–47, 50, 51, 54, 55, 57, 60, 61] recently from both academia and industry. The benefit of mixing NVM with DRAM for Big Data systems is obvious — NVM’s
high capacity makes it possible to fulfill the high memory requirement of a Big Data workload with a small number of compute nodes, holding the promise of significantly reducing the costs of both hardware and energy in large data centers.

1.1 Problems

Although using NVM for Big Data systems is a promising direction, the idea has not yet been fully explored. Adding NVM naively would lead to large performance penalties due to its significantly increased access latency and reduced bandwidth — e.g., the latency of an NVM read is 2–4× larger than that of a DRAM read and NVM’s bandwidth is about 1/8 - 1/3 of that of DRAM [19, 49]. Hence, a critical research question that centers around all hybrid-memory-related research is how to perform intelligent data allocation and migration between DRAM and NVM so that we can maximize the overall energy efficiency while minimizing the performance overhead? To answer this question in the context of Big Data processing, there are two major challenges.

**Challenge #1: Working with Garbage Collection (GC).** A common approach to managing hybrid memories is to modify the OS or hardware to (1) monitor access frequency of physical memory pages, and (2) move the hot (frequently-accessed) data into DRAM. This approach works well for native language applications where data stays in the memory location it is allocated into. However, in managed languages, the garbage collector keeps changing the data layout in memory by copying objects to different physical memory pages, which breaks the bonding between data and physical memory address. Most Big Data systems are written in such managed languages, e.g., Java and Scala, for the quick development cycle and rich community support they provide. Managed languages are executed on top of a managed runtime such as the JVM, which employs a set of sophisticated memory management techniques such as garbage collection. As a traditional garbage collector is not aware of hybrid memories, allocating and migrating hot/cold pages at the OS level can easily lead to interference between these two different levels of memory management.

**Challenge #2: Working with Application-level Memory Subsystems.** Modern Big Data systems all contain sophisticated memory subsystems that perform various memory management tasks at the application level. For instance, Apache Spark [5] uses resilient distributed datasets (RDDs) as its data abstraction. An RDD is a distributed data structure that is partitioned across different servers. At a low level, each RDD partition is an array of Java objects, each representing a data tuple. RDDs are often immutable but can exhibit diverse lifetime behavior. For example, developers can explicitly persist RDDs in memory for memoization or fault tolerance. Such RDDs are long-lived while RDDs storing intermediate results are short-lived.

An RDD can be at one of many storage levels (e.g., memory, disk, unmaterialized, etc.). Spark further allows developers to specify, with annotations, where an RDD should be allocated, e.g., in the managed heap or native memory. Objects allocated natively are not subject to GC, leading to increased efficiency. However, data processing tasks, such as shuffle, join, map, or reduce, are performed over the managed heap. A native-memory-based RDD cannot be directly processed unless it is first moved into the heap. Hence, where to allocate an RDD depends on when and how it is processed. For example, a frequently-accessed RDD should be placed in DRAM while a native-memory-based RDD would not be frequently used and placing it in NVM would be desirable. Clearly, efficiently using hybrid memories requires appropriate coordination between these orthogonal data placement policies, i.e., the heap, native memory, or disk, vs. NVM or DRAM.

**State of the Art.** In summary, the key challenges in supporting hybrid memories for Big Data processing lie in how to develop runtime system techniques that can make memory allocation/migration decisions that match how data is actually used in an application. Although techniques such as Espresso [54] and Write Rationing [9] support NVM for managed programs, neither of them was designed for Big Data processing whose data usage is greatly different than that of regular, non-data-intensive Java applications [41, 42].

For example, Espresso defines a new programming model that can be used by the developer to allocate objects in persistent memory. However, real-world developers would be reluctant to completely re-implement their systems from scratch using such a new model. Shoaib et al. [9] introduced the Write Rationing GC, which moves the objects that experience a large/small number of writes into DRAM/NVM to prolong NVM’s lifetime. Write Rationing pioneers the work of using the GC to migrate objects based on their access patterns. However, Big Data systems make heavy use of immutable datasets — for example, in Spark, most RDDs are immutable. Placing all immutable RDDs into NVM can incur a large overhead as many of these RDDs are frequently read and an NVM read is 2–4× slower than a DRAM read.

1.2 Our Contributions

**Our Insight.** Big Data applications have two unique characteristics that can greatly aid hybrid memory management. First, they perform bulk object creation, and data objects exhibit strong epochal behavior and clear access patterns. For example, Spark developers program with RDDs, each of which contains objects with exactly the same access/lifetime patterns. Exploiting these patterns at the runtime would make it much easier for Big Data applications to enjoy the benefits of hybrid memory.

Second, the data access and lifetime patterns are often statically observable in the user program. For example, an RDD is a coarse-grained data abstraction in Spark and the
access patterns of different RDDs can often be inferred from the way they are created and used in the program (§2).

Hence, unlike regular, non-data-intensive applications for which profiling is often needed to understand the access patterns of individual objects, we can develop a simple static analysis for a Big Data application to infer the access pattern of each coarse-grained data collection, in which all objects share the same pattern. This observation aligns well with prior work (e.g., Façade [42] or Yak [41]) that requires simple annotations to specify epochs to perform efficient garbage collection for Big Data systems. The static analysis does not incur any runtime overhead, yet it can produce precise enough data access information for the runtime system to perform effective allocation and migration.

**Panthera.** Based on our extensive experience with Big Data applications, we propose Panthera, which divides a mess of data objects into several data collections according to application’s semantics and infers the coarse-grained data usage behavior by light-weight static program analysis and dynamic data usage monitoring. Panthera leverages garbage collection to migrate data between DRAM and NVM, incurring almost no runtime overhead.

We focus on Apache Spark in this paper as it is the de-facto data-parallel framework deployed widely in industry. Spark hosts a range of applications in machine learning, graph analytics, stream processing, etc., making it worthwhile to build a specialized runtime system, which can provide immediate benefit to all applications running atop. Although Panthera was built for Spark, our idea is applicable to other systems such as Hadoop as well; §4 provides a detailed discussion of Panthera’s applicability.

Panthera enhances both the JVM and Spark with two major innovations. First, based on the observation that access patterns in a Big Data application can be identified statically, we develop a static analysis (§3) that analyzes a Spark program to infer a memory tag (i.e., NVM or DRAM) for each RDD variable based on the variable’s location and the way it is used in the program. These tags indicate which memory the RDDs should be allocated in.

Second, we develop a new semantics-aware and physical-memory-aware generational GC (§4). Our static analysis instruments the Spark program to pass the inferred memory tags down to the runtime system, which uses these tags to make allocation/migration decisions. Since our GC is based on a high-performance generational GC in OpenJDK, Panthera’s heap has two spaces, representing a young and an old generation. We place the entire young generation in DRAM while splitting the old generation into a small DRAM component and a large NVM component. The insight driving this design is based on a set of key observations (discussed in §2 in detail) we make over the lifetimes and access patterns of RDDs in representative Spark executions:

- Most objects are allocated initially in the young generation. Since they are frequently accessed during initialization, placing them in DRAM enables fast access to them.
- Long-lived objects in Spark can be roughly classified into two categories: (1) long-lived RDDs that are frequently accessed during data transformation (e.g., cached for iterative algorithms) and (2) long-lived RDDs that are cached primarily for fault tolerance. The first category of RDDs should be placed in the DRAM component of the old generation because they have long lifespans and DRAM provides desirable performance for frequent access to them. The second category should be placed in the NVM component of the old generation because they are infrequently accessed and hence NVM’s large access latency has relatively small impact on overall performance.
- There are also short-lived RDDs that store temporary, intermediate results. These RDDs die and are then reclaimed in the young generation quickly, leading to frequent accesses to this area. This is another reason why we place the young generation within DRAM.

Based on these observations, we modified both the minor and major GC, which allocate and migrate data objects, based on their RDD types and the semantic information inferred by our static analysis, into the spaces that best fit their lifetimes and access patterns. Our runtime system also monitors the transformations invoked over RDD objects to perform runtime (re)assessment of RDDs’ access patterns. Even if the static analysis does not accurately predict an RDD’s access pattern and the RDD gets allocated in an undesirable space, Panthera can still migrate the RDD from one space to another using the major GC.

**Results.** We have evaluated Panthera extensively with graph computing (GraphX), machine learning (MLlib) and other iterative in-memory computing applications (Table 4) Results with various heap sizes and DRAM ratios demonstrate that Panthera makes effective use of hybrid memories — overall, the Panthera-enhanced JVM reduces the memory energy by 37%–52% with only a 1%–4% execution time overhead, whereas Write Rationing [9] that moves read-only RDD objects into NVM incurs a 41% time overhead.

2 Background and Motivation

This section provides necessary background on Apache Spark [5] and a motivating example that illustrates the access patterns in a Spark program.

**Spark Basics.** Spark is a data-parallel system that supports acyclic data flow and in-memory computing. The major data representation used by Spark is resilient distributed dataset (RDD) [58], which represents a read-only collection of tuples. An RDD is a distributed memory abstraction partitioned in the cluster. Each partition is an array of data items of the same type. Each node maintains an RDD partition, which is essentially a multi-layer Java data structure — a top RDD object references a Java array, which, in turn, references a
A narrow dependence exists from a parent to a child RDD if and executed in parallel. Transformations that exhibit narrow dependences are grouped into the same stage starts by reading data from disk files. However, for wide dependences, shuffling is necessary. For example, for narrow dependences makes it possible for Spark to determine whether a child partitions. Distinguishing these two types of dependences is classified into “narrow” and “wide”.

Wide dependences are characterized by the fact that each partition of the parent is used by at most one partition of the child. A Spark pipeline consists of a sequence of transformations and actions over RDDs. A transformation produces a new RDD from a set of existing RDDs; examples are map, reduce, or join. An action is a function that computes statistics from an RDD, such as an aggregation. Spark leverages lazy evaluation for efficiency, that is, a transformation may not be evaluated until an action is performed later on the resulting RDD. Before data processing starts, the dependences between RDDs are first extracted from the transformations to form a lineage graph, which can be used to conduct lazy evaluation and RDD recomputation upon node failures.

With lazy evaluation, a transformation only creates a (top-level) RDD object without materializing the RDD (i.e., the point at which its internal array and actual data tuples are created). Recomputing all RDDs is time-consuming when the lineage is long or when it branches out, and hence, Spark allows developers to cache certain RDDs in memory (by using the API persist). Developers can specify a storage level for a persisted RDD, e.g., in memory or on disk, in the serialized or deserialized form, etc. RDDs that are not explicitly persisted are temporary RDDs that will be garbage-collected when they are no longer used, while persisted RDDs are materialized and never collected.

The Spark scheduler examines the lineage graph to build a DAG of stages for execution. The lineage (transformation)-based dependences are classified into “narrow” and “wide”. A narrow dependence exists from a parent to a child RDD if each partition of the parent is used by at most one partition of the child RDD. By contrast, a wide dependence exists when each partition of the parent RDD may be used by multiple child partitions. Distinguishing these two types of dependences makes it possible for Spark to determine whether a shuffle is necessary. For example, for narrow dependences, shuffling is not necessary, while for wide dependences it is.

A Spark pipeline is split into a set of stages based on shuffles (and thus wide dependences) — each stage ends at a shuffle that writes RDDs onto the disk and the next stage starts by reading data from disk files. Transformations that exhibit narrow dependences are grouped into the same stage and executed in parallel.

**RDD Characteristics.** An RDD is, at a low level, an array of Java objects, which are managed by the semantics-agnostic GC in the JVM. RDDs often exhibit predictable lifetime and memory-access patterns. Our goal is to pass these patterns down to the GC, which can exploit such semantic information for efficient data placement. We provide a concrete example to illustrate these patterns.

Figure 2(a) shows the Spark program for PageRank [14], which is a well-known graph algorithm used widely by search engines to rank web pages. The program iteratively computes the rank of each vertex based on the contributions of its in-neighbors. Three RDDs can be seen from its source code: links representing edges from the input graph, contribs containing contributions from incoming edges of each vertex, and ranks that maps each vertex to its page rank. links is a static map computed from the input while contribs and ranks are recomputed per iteration of the loop.

In addition to these three developer-defined RDDs visible in the program, Spark generates many invisible RDDs to store intermediate results during execution. A special type of intermediate RDD is ShuffledRDD. Each iteration of the loop in the example forms a stage that ends at a shuffle, writing shuffled data into different disk files. In the beginning of the next stage, Spark creates a ShuffledRDD as input for the stage. Unlike other intermediate RDDs that are never materialized, ShuffledRDDs are immediately materialized because they contain data read freshly out of disk files. However, since they are not persisted, they will be collected when the stage is completed.

In summary, (1) persisted RDDs are materialized at the moment the method persist is called, and (2) non-persisted RDDs are not materialized unless they are ShuffledRDDs or an action is invoked on them.

**Example.** Figure 2(b) shows the set of RDDs that exists within a stage (i.e., iteration) and their dependences. Suppose each RDD has three partitions (on three nodes). The dashed edges represent wide dependences (i.e., shuffles) due to the reduction on Line 17. There are totally eight RDDs generated in each iteration. ShuffledRDD[8], which stems from the reduction on Line 17, is transformed to ranks via a map transformation. ranks joins with links to form CoGroupedRDD[3], which is then processed by four consecutive map functions, i.e., \( f_4 \circ \ldots \circ f_1 \), producing contribs at the end. For unmaterialized (blue) RDDs, the sequence of transformations (e.g., \( f_4 \circ \ldots \circ f_1 \)) is applied to each record from the source RDD in a streaming manner via iterators to produce a final record.

For links and contribs, the developer invokes the method persist to materialize these RDDs. The storage levels indicate that links is cached in memory throughout the execution (as it is used in each iteration) while contribs generated in each iteration is kept in memory but will be serialized to disk upon memory pressure. ranks is not explicitly persisted. Hence, it is not materialized until the execution reaches Line 20 where action count is invoked on the RDD object.
Panthera: Holistic Memory Management for Big Data ...

PLDI ’19, June 22–26, 2019, Phoenix, AZ, USA

The lifetime patterns of these different RDDs fall into two categories. Non-persisted intermediate RDDs are short-lived as their data objects are generated only during a pipelined execution. Persisted RDDs are long-lived and stay in memory/on disk until the end of the execution. Their access patterns are, however, more diverse. Objects in an intermediate RDD are accessed at most once during streaming. Objects in a persisted RDD can exhibit different types of behavior. For RDDs like `links` that are used in each iteration, their objects are frequently accessed. In contrast, RDDs like `contribs` are persisted primarily for speeding up recovery from faults, and hence, their objects are rarely used after generated.

**Design Choices.** The different characteristics of DRAM and NVM make them suitable for different types of datasets. DRAM has low capacity and fast access speed, while NVM has large capacity but slow speed. Hence, DRAM is a good choice for storing small-sized, frequently-accessed datasets, while large-sized, infrequently-accessed datasets fit naturally into NVM. The clear distinction in the lifespans and access patterns of different RDDs makes it easy for them to be placed into different memories suitable for their behavior. For example, intermediate (blue) RDDs are never materialized. Their objects are created individually during streaming and then quickly collected by the GC. These objects are allocated in the young generation and will eventually die there. Although they are short-lived, they are accessed frequently during streaming. This motivates our design choice of placing the young generation in DRAM.

Persisted RDDs, in contrast, have all their data objects created at the same time, and thus need large storage space. Since they are kept alive indefinitely, they should be allocated directly in the old generation. One category of persisted RDDs includes those that are frequently accessed, like `links`; they need to be placed in DRAM. Another category includes RDDs that are rarely accessed and cached for fault tolerance, like `contribs`, these RDDs should be placed in NVM. This behavioral difference motivates our choice of splitting the old generation into a DRAM and an NVM component.

We perform what we suggest on a system with 32GB DRAM and 88GB NVM using Spark-based PageRank as the benchmark. Figure 2(c) shows the performance and energy consumption normalized to a system with 120GB of DRAM. Compared to using only 32GB DRAM, adding 88GB NVM to the system provides modest performance benefit (15%) but leads to 47% higher energy consumption, without proper data placement across DRAM and NVM (see Unmanaged, §5.2). After applying Panthera, RDD `links` and `contribs` are placed into DRAM and NVM, respectively. With such careful placement of data across DRAM and NVM, we find that (1) performance increases by 42% compared to using only a 32GB DRAM, and becomes at the same level of the performance of using 120GB DRAM, (2) energy consumption is only 9% more than using only a 32GB DRAM, and 40% better than using a 120GB DRAM. We conclude that careful data placement between DRAM and NVM can provide the performance of large DRAM system, while keeping the energy consumption at the level of a small DRAM system.

### 3 Static Inference of Memory Tags

By analyzing the access patterns of RDDs, we can determine their behavior and allocate them to the appropriate memory. This information is then used by the runtime system to make efficient data allocation decisions.

```
var lines = ctx.textFile(args[0]).slices();
var parts = s.split("\s+");
var contribs = links.join(ranks).values().flatMap{
  case(urls, rank) =>
    val size = urls.size
    val map = urls.map(_.url => (url, rank / size))
    .persist(StorageLevel.MEMORY_AND_DISK_SER)
    .mapValues(0.15 + 0.85 * _)}
    .reduceByKey(_ + _)
    .mapValues(0.15 + 0.85 * _)
  }
ranks.count();
```

![Figure 2. Characteristics of RDDs in Spark PageRank](image)
**Static Analysis.** In a Spark program, the developer can invoke `persist` with a particular storage level on an RDD to materialize the RDD, as illustrated in Figure 2. We piggyback on the storage levels to further determine if a persisted RDD should be placed into DRAM or NVM. In particular, Panthera statically analyzes the program to infer a memory tag (i.e., DRAM or NVM) for each persist call. Each of the ten existing storage levels (e.g., `MEMORY_ONLY`), except for `OFF_HEAP` and `DISK_ONLY`, is expanded into two sub-levels, annotated with NVM and DRAM, respectively (e.g., `MEMORY_ONLY_DRAM` and `MEMORY_ONLY_NVM`). `OFF_HEAP` is translated directly into `OFF_HEAP_NVM` because RDDs placed in native memory are rarely used, while `DISK_ONLY` does not carry any memory tag.

Our static analysis performs inference based on the `def-use` information w.r.t. each RDD variable declared in the program as well as the loop(s) in which the variable is defined/used. Our key insight is that if the variable is defined in each iteration of a computational loop, most of the RDD instances represented by the variable are not used frequently. This is because Spark RDDs are often immutable and hence, every definition of the RDD variable creates a new RDD instance at run time, leaving the old RDD instance cached and unused. Hence, we tag the variable “NVM”, instructing the runtime system to place these RDDs in NVM. An example is the `contribs` variable in Figure 2(a), which is defined in every iteration of the loop — although the variable is also used in each iteration, the use refers to the most recent RDD instance created in the last iteration while the instances created in all the other past iterations are left unused.

By contrast, if a variable is used-only (i.e., never defined) in the loop, such as `links`, we create a tag “DRAM” for it since only one instance of the RDD exists and is repeatedly used. Panthera analyzes not only RDD variables on which `persist` is explicitly called, but also those on which actions are invoked, such as the `ranks` variable in Figure 2(a). The tag inferred for an RDD variable (say v) is passed, at the materialization point of every RDD instance (v refers to), into the runtime system via automatically instrumented calls to auxiliary (native) methods provided by the Pantera JVM. We piggyback on a tracing GC to propagate this tag from the RDD object down to each data object contained in the RDD — when the GC runs, it moves objects with the same tag together into the same (DRAM or NVM) region (see §4).

One constraint that needs to be additionally considered is the location of the loop relative to the location of the materialization point of the RDD. We analyze the loop only if the materialization point `precedes or is in` the loop. Otherwise, whether the variable is used or defined in the loop does not matter as the RDD has not been materialized yet. For instance, although the `ranks` variable is defined in the loop that starts at Line 17, it does not get materialized until Line 20 after the loop finishes. Hence, its behavior in the loop does not affect its memory tag, which should actually depend on its `def-use` in the loops, if any, after Line 20.

If no loop exists in a program, the program has only one iteration and all RDDs receive an “NVM” tag as none of them are repeatedly accessed. If there are multiple loops to be considered for an RDD variable, we tag it “DRAM” if there exists one loop in which the variable is used-only and that loop follows or contains the materialization point of the RDD. The variable receives an “NVM” tag otherwise. If all persisted RDDs receive an "NVM" tag at the end of the analysis, we change the tags of all RDDs to "DRAM" — the goal is to fully utilize DRAM by first placing RDDs in DRAM. Once DRAM capacity is exhausted, the remaining RDDs, including those with a “DRAM” tag, will be placed in NVM.

Note that our analysis infers tags only for the RDD variables explicitly declared in the program. Intermediate RDDs produced during execution are not materialized and thus do not receive memory tags from our analysis. We discuss how to handle them in §4.

The memory tag of an RDD variable is a static approximation of its access pattern, which may not reflect the behavior of all RDD instances represented by the variable at run time. However, user code for data processing often has a simple batch-transformation logic. Hence, the static information inferred from our analysis is often good enough to help the runtime make an accurate placement decision for the RDD. In case the statically inferred tags do not precisely capture the RDD’s access information, Pantera has the ability to move RDDs between NVM and DRAM (within the old generation) based on their access frequencies, when a full-heap GC occurs. §4 provides a full discussion for this mechanism.

**Dealing with ShuffledRDD.** Recall from §2 that, in addition to the RDDs on which `persist` is explicitly invoked, `ShuffledRDDs`, which are created from disk files after a shuffle, are also materialized. These RDDs are often the input of a stage but invisible in the program code. The challenge here is where to place them. Our insight is that their placement should depend on the other materialized RDDs that are transformed from (i.e., depend on) them in the same stage.

For example, in Figure 2(b), the input of the stage are two sets of `ShuffledRDDs`: [1] and [8]. `ShuffledRDD[1]` is the RDD represented by `links` and our static analysis already infers tag “DRAM” for it. `ShuffledRDD[8]` results from the reduction in the previous stage. Because `ShuffledRDD[8]` transitive produces `MapPartitionRDD[7]` (represented by `contribs`) and `MapPartitionRDD[7]` has a memory tag “NVM” inferred by our static analysis, we tag `ShuffledRDD[8]” NVM” as well.

The main reason that RDDs belonging to the same stage may share many data objects for optimization purposes. For example, a map transformation that only changes the values (of key-value pairs) in RDD A may generate a new RDD B that references the same set of key objects as in A. If B has already received a memory tag from our static analysis, it is better to assign the same tag to A so that these shared objects do not receive inconsistent tags and would not need to be moved from one memory to another when B is generated.
memory is placed entirely in NVM. We reserve two unused bits, referred to as MEMORY_BITS, from the header of each object to indicate whether the object should be allocated into DRAM (01) or NVM (10). The default value for these bits is 00 — objects that do not receive a tag have this default value. They will be promoted to the NVM component of the old generation if they live long enough. Figure 3 illustrates the heap structure and our allocation policies.

**Allocation Policies.** As discussed in §3, each materialized RDD carries a memory tag that comes from our static analysis or lineage-based tag propagation. However, at a low level, an RDD is a structure of objects, as illustrated in Figure 1, and these objects are created at different points in the execution. Our goal is to place all objects belonging to the same logical RDD — including the top object, the array object, tuple objects, and other objects reachable from tuples — together in the space suggested by the RDD’s memory tag, because these objects likely have the same access pattern and lifetime.

However, this is rather challenging — our static analysis infers a memory tag for each top RDD object (whose type is a subtype of org.apache.spark.rdd.RDD) in the user program and we do not know what other objects belong to this RDD by just analyzing the user program. Statically identifying what objects belong to a logical data structure would require precise context-sensitive static analysis of both user and system code, which is difficult to do due to Spark’s extremely large codebase and the scalability issues of static analysis.

Our idea to solve this problem is that instead of attempting to allocate all objects of an RDD directly into the space (say S) suggested by the RDD’s tag, we allocate only the array object into S upon its creation. This is much easier to do — Panthera instruments each materialization point (e.g., before a call to persist or a Spark action) in the user program to pass the tag down to the runtime system without needing to analyze the Spark system code. Since the array is created at materialization, the runtime system can just use the tag to determine where to place it. All other objects in the RDD are not immediately allocated in S due to the aforementioned difficulties in finding their allocation sites. They are instead allocated in the young generation. Later, we use the GC to move these objects into S as tracing is performed.

Another important reason why we first allocate the array object into S is because the array is often much larger than the top and tuple objects. It is much more efficient to allocate it directly into the space it belongs to rather than allocating it somewhere else and moving it later.

Table 1 shows our allocation policies for different types of objects in an RDD. For RDDs with tag “DRAM”, array objects are allocated directly into the DRAM component of the old generation if it has enough space. Otherwise, they have to be allocated in the NVM component. For RDDs with tag “NVM”, array objects are allocated directly into the NVM component. Intermediate RDDs without tags are all allocated
in the young generation (DRAM). Most of them end up dying there and never get promoted, while a small number of objects that eventually become old enough will be promoted to the NVM space of the old generation. Top RDD objects and data tuple objects, as discussed earlier, are all allocated into the young generation and moved later by the GC to the spaces containing their corresponding arrays.

Table 1. Panthera’s allocation policies.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Obj Type</th>
<th>Initial Space</th>
<th>Final Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>RDD Top</td>
<td>Young Gen.</td>
<td>DRAM of Old Gen.</td>
</tr>
<tr>
<td></td>
<td>Data Obj</td>
<td>Young Gen.</td>
<td>DRAM of Old Gen.</td>
</tr>
<tr>
<td>NVM</td>
<td>RDD Top</td>
<td>Young Gen.</td>
<td>NVM of Old Gen.</td>
</tr>
<tr>
<td></td>
<td>Data Obj</td>
<td>Young Gen.</td>
<td>NVM of Old Gen.</td>
</tr>
<tr>
<td>NONE</td>
<td>RDD Array</td>
<td>Young Gen.</td>
<td>Young Gen. or NVM of Old Gen.</td>
</tr>
<tr>
<td></td>
<td>Data Obj</td>
<td>Young Gen.</td>
<td>Young Gen. or NVM of Old Gen.</td>
</tr>
</tbody>
</table>

4.2 Implementation and Optimization

This subsection describes our implementation techniques and various optimizations.

4.2.1 Passing Tags

Right before each materialization point (i.e., the invocation of persist or a Spark action), our analysis inserts a call to a native method `rdd_alloc(rdd, tag)`, with the RDD’s top object (rdd) and the inferred memory tag (tag) as the arguments. This method first sets a thread-local state variable to DRAM or NVM, according to the tag, informing the current thread that a large array for an RDD will be allocated soon. Next, `rdd_alloc` sets the `MEMORY_BITS` of the top object rdd based on tag. Regardless of where it currently is, this top object will eventually be moved by the GC to the space corresponding to tag.

The thread then transitions into a “wait” state, waiting for this large array. In this state, the first allocation request for an array whose length exceeds a user-defined threshold (i.e., a million used in our experiments) is recognized as the RDD array. Panthera then allocates the array directly into the space indicated by tag. To implement this, we modified both the fast allocation path, assembly code generated by the JIT compiler, and the slow path, functions implemented in C++. After this allocation, the state variable is reset and the thread exits the wait state. If tag is null, the array is allocated in the young generation, preferably through the thread-local allocation buffer (TLAB), and the `MEMORY_BITS` of the top object remains as the default value (00).

4.2.2 Object Migration

There are two major challenges in how to move objects: cross-generation migration and object compaction. As Panthera piggybacks on a generational GC, objects in the young generation that survive several minor GCs are deemed long-lived and moved into the old generation. We leverage this opportunity to move together objects that belong to the same logical RDD — as discussed earlier, these objects might not have been allocated in the same space initially.

Minor GC. To do this, we modified the minor collection algorithm in the Parallel Scavenge GC on which Panthera is built. The existing minor GC contains three tasks: root-task, which performs object tracing from the roots (e.g., stack and global variables); old-to-young-task, which scans references from objects in the old generation to those in the young generation to identify (directly or transitively) reachable objects; and steal-task, which performs work stealing for load balancing. To support our object migration, we split old-to-young-task into a DRAM-to-young-task and NVM-to-young-task, which find objects that should be moved into the DRAM and NVM parts of the old generation, respectively.

For these two tasks, we modified the tracing algorithm to propagate the tag — for example, scanning a reference from a DRAM-based RDD array (with tag “DRAM”) to a tuple object (in the young generation) propagates the tag to the tuple object (by setting its `MEMORY_BITS`). Hence, when tracing is done, all objects reachable from the array have their `MEMORY_BITS` set to the same value as that of the array. In the original GC algorithm, an object does not get promoted from the young to the old generation until it survives several minor GCs. In Panthera, however, we move the objects whose `MEMORY_BITS` is set as 01 (10) in tracing immediately to DRAM (NVM) space in the old generation. We refer to this mechanism as eager promotion. Objects whose `MEMORY_BITS` is not set, 00, in tracing belong to intermediate RDDs or are control objects not associated with any RDDs. The migration of these objects follows the original algorithm, that is, they will be moved only if they survive several minor GCs.

Furthermore, we also need to move RDD top objects to the appropriate part of the old generation. These top objects, whose `MEMORY_BITS` was set by the instrumented call to `rdd_alloc` at their materialization points, are visited when root-task is executed because these objects are referenced directly by stack variables. We modified the root-task algorithm to identify objects with the set `MEMORY_BITS`. These RDD top objects will also be moved to (the DRAM (01) or NVM (10) space of) the old generation by the minor GC.

Major GC. When a major GC runs, it performs memory compaction by moving objects together (in the old generation) to reduce fragmentation and improve locality. We modified the major GC to guarantee that compaction does not occur across the boundary between DRAM and NVM. Furthermore, when the major GC performs a full-heap scan, Panthera re-assesses, for each RDD array object, where the object should actually be placed based on the RDD’s runtime access frequency. This frequency is measured by counting, using instrumentation, how many times a method (e.g., `map` or `reduce`) has been invoked on this RDD object.

We maintain a hash table that maps each RDD object to the number of calls made on the object. Our static analysis inserts, at each such call site, a JNI (Java Native Interface) call that invokes a native JVM method to increment the call
frequency for the RDD object. Frequently (infrequently) accessed array objects are moved from the NVM (DRAM) space to the DRAM (NVM) space within the old generation and all objects reachable from these arrays are moved as well. Their MEMORY_BITS will be updated accordingly. At the end of each major GC, the frequency for each RDD is reset.

The DRAM space of the old generation can be quickly filled up as it is much smaller than the NVM space. When the DRAM space is full, the minor GC moves all objects from the young generation to the NVM space of the old generation regardless of their memory tags.

**Conflicts.** If an object is reachable from multiple references and different tags are propagated through them, a conflict occurs. As discussed earlier, we resolve conflicts by giving “DRAM” higher priority than “NVM”. As long as the object receives “DRAM” from any reference, it is a DRAM object and will be moved to the DRAM space of the old generation.

### 4.2.3 Card Optimization

In OpenJDK, the heap is divided into many cards, each representing a region of 512 bytes. Every object can take one or more cards, and the write barrier maintains a card table that marks certain cards dirty upon reference writes. The card table can be used to efficiently identify references during tracing. For example, upon \( a.f = b \), the card that contains the object referenced by \( a \) is set to dirty. When a minor GC runs, the old-to-young scavenger task cleans a card if the target objects of the (old-to-young) references contained in the memory region represented by the card have been copied to the old generation.

However, if a card contains two large arrays (say \( A \) and \( B \)) — \( e.g. \), \( A \) ends in the middle of the card while \( B \) starts there immediately — significant inefficiencies can result when they are scanned by two different GC threads. The card would remain dirty even if all objects referenced by \( A \) and \( B \) have been moved from the young to the old generation — neither thread could clean the card due to its unawareness of the status of the array scanned by another thread. This would cause every minor GC to scan every element of each array in the dirty card until a major GC occurs.

This is a serious problem for Big Data applications that make heavy use of large arrays. Shared cards exist pervasively when these arrays are frequently allocated and deallocated. Frequent scanning of such cards with multiple threads can incur a large overhead on NVM due to its higher read latency and reduced bandwidth. We implemented a simple optimization that adds an alignment padding for the allocation of each RDD array to make the end of the array align with the end of a card. Although this leads to space inefficiencies, the amount of wasted space is small (\( e.g. \), less than 512 bytes for each array of hundreds of megabytes) while card sharing among arrays is completely eliminated, resulting in substantial reduction in GC time.

### 4.3 Applicability

Our static analysis is designed specifically for Spark and not easily reusable to other framework. However, the APIs for data placement and migration provided by the Panthera runtime system can be employed to manage memory for any Big Data system that uses a key-value array as its backbone data structure. Examples include Apache Hadoop, Apache Flink, or database systems such as Apache Cassandra.

Panthera provides two major APIs, one for pre-tenuring data structures with tags and a second for dynamic monitoring and migration. The first API takes as input an array and a tag, performing data placement as discussed earlier in this section. The tag can come from the developer’s annotations in the program or from a static analysis that is designed specifically for the system to be optimized.

To illustrate, consider Apache Hadoop where both a map worker and a reduce worker may need to hold large data structures in memory. Some of these data structures are loaded from HDFS as immutable input, while others are frequently accessed. In the case of HashJoin, which is a building block for SQL engines, one input table is loaded entirely in memory while the second table is partitioned across map workers. If map workers are executed in separate threads, they all share the first table and join their own partitions of the second table with it. The first table is long-lived and frequently accessed. Hence, it should be tagged DRAM and placed in the DRAM space of the old generation, while different partitions of the second table can be placed in the young generation and they will die there quickly.

Panthera’s second API takes as input a data structure object to track the number of calls made on the object. If this API is used to track the access frequency of the data structure, the data structure (and all objects reachable from it) would not be pretenured (as specified by the first API), but rather, they are subject to dynamic migration performed in the major GC. We can use this API to dynamically monitor certain objects and migrate them if their access patterns are not easy to predict statically.

Use of these two APIs enables a flexible allocation/migration mechanism that allows certain parts of the data structure (\( e.g., \) for which memory tags can be easily inferred) to be pretenured and other parts to be dynamically migrated.

### 5 Evaluation

We have added/modified 9186 lines of C++ code in OpenJDK (build jdk8u76-b02) to implement the Panthera GC and written 979 lines of Scala code to implement the static analysis.

#### 5.1 NVM Emulation and Hardware Platform

Most of the prior works on hybrid memories used simulators for experiments. However, none of them support Java applications well. We cannot execute managed-runtime-based
distributed systems on these simulators. There also exist emulators such as Quartz [49] and PMEP [18] that support emulation of NVM for large programs using commodity multi-socket (NUMA) hardware, but neither Quartz nor PMEP could run OpenJDK. These emulators require developers to use their own libraries for NVM allocation, making it impossible for the Panthera GC to migrate objects without re-implementing the entire allocator and GC from scratch using these libraries.

As observed in [8, 49], NUMA’s remote memory latency is close to NVM’s latency, and hence, researchers have used a NUMA architecture as the baseline to measure emulation accuracy. Following this observation, we built our own emulator on NUMA machines to emulate hybrid memories for JVM-based Big Data systems. We followed Quartz [49] when implementing our emulator. Quartz has two major components: (1) it uses the thermal control register to limit the DRAM bandwidth; and (2) it creates a daemon thread for each application process and inserts delay instructions to emulate the NVM latency. For example, if an application’s CPU stall time is \( S \), Quartz scales its latency to \( S \times \frac{NVM\_latency}{DRAM\_latency} \) to emulate the effect of NVM. For (1), we used the same thermal control register to limit the read/write bandwidth. Like Quartz, we currently do not support different bandwidths for reads and writes. For (2), we followed Quartz’s observation to use the latency of NUMA’s remote memory to model NVM’s latency.

An alternative approach to emulating NVM’s latency is to instrument loads/stores during JIT compilation, injecting a software-created delay at each load/store. The limitation of this approach, however, is that it does not account for caching effects and memory-level parallelism.

We used one CPU to run all the computation, the memory local to the CPU as DRAM, and the remote memory as NVM. In particular, DRAM and NVM are emulated, respectively, using 2 local and 2 remote memory channels. The performance specifications of the emulated NVM are the same as those used in [49], reported in Table 2. We obtained the NVM configurations from Micron’s specification [38]. To emulate NVM’s slow write speed, we used the thermal control register to limit the read/write bandwidth. Like Quartz, we currently do not support different bandwidths for reads and writes. For (2), we followed Quartz’s observation to use the latency of NUMA’s remote memory to model NVM’s latency.

An alternative approach to emulating NVM’s latency is to instrument loads/stores during JIT compilation, injecting a software-created delay at each load/store. The limitation of this approach, however, is that it does not account for caching effects and memory-level parallelism.

We used one CPU to run all the computation, the memory local to the CPU as DRAM, and the remote memory as NVM. In particular, DRAM and NVM are emulated, respectively, using 2 local and 2 remote memory channels. The performance specifications of the emulated NVM are the same as those used in [49], reported in Table 2. We obtained the NVM configurations from Micron’s specification [38]. To emulate NVM’s slow write speed, we used the thermal control register to limit the read/write bandwidth. Like Quartz, we currently do not support different bandwidths for reads and writes. For (2), we followed Quartz’s observation to use the latency of NUMA’s remote memory to model NVM’s latency.

Energy Estimation. We followed Lee et al. [31] to estimate energy for NVM. We used Micron’s DDR4 device specifications [38] to model DRAM’s power. NVM’s energy has a static and dynamic component. The static component is negligible compared to DRAM [32]. The dynamic component consists of the energy consumed by reads and writes. Reads on NVM consume less energy than on DRAM as they do not require restoration due to their non-destructive nature.

NVM writes consume more energy than DRAM writes. Upon a row-buffer miss, the energy consumed by each write has three components: (1) an array write that evicts data from the row buffer into the bank array, (2) an array read that fetches data from the bank array to the row buffer, and (3) a row buffer write that writes new data from the CPU last level cache to the row buffer. Assuming the row-buffer miss ratio is 0.5, we computed these three components separately by considering the row buffer’s write energy (1.02pJ/bit), size (i.e., 8K bits for DRAM [38], 32-bit-wide partial writes back for NVM [31]) and miss rate (0.5), as well as the array’s write-back energy (16.8pJ/bit × 7.6% for NVM) and read energy (2.47pJ/bit for NVM). The factor of 7.6% is due to Lee et al.’s optimization [31] that writes only 7.6% of the dirty words back to the NVM array.

CPU’s uncore events, collected with VTune [6], were employed to compute the numbers of reads and writes. In particular, the events we used were \( \text{UNC}_\text{M\_CAS\_COUNT.RD} \) and \( \text{UNC}_\text{M\_CAS\_COUNT.WR} \). VTune can also distinguish reads and writes from/to local and remote memories.

5.2 Experiment Setup

We set up a small cluster to run Spark with one master node and one slave node — these two servers have a special Intel chipset with a “scalable memory buffer” that can be tuned to produce the 2.6× latency for remote memory accesses, which matches NVM’s read/write latency. Since our focus is not on distributed computing, this small cluster is sufficient for us to execute real workloads on Spark and understand their performance over hybrid memories. Table 3 reports the hardware configurations of the Spark master and Spark slave nodes. Each node has an 8-core CPU and the Parallel Scavenge collector on which Panthera was built creates 16 GC threads in each (minor or major) GC to perform parallel tracing and compaction.

Table 3. Hardware configuration for both servers.

<table>
<thead>
<tr>
<th>Arch</th>
<th>NUMA, 4 sockets</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>E7-4809 v3 2.00GHz, 8 cores, 16 HW threads</td>
</tr>
<tr>
<td>L1-I</td>
<td>8 way, 32KB/core, private</td>
</tr>
<tr>
<td>L1-D</td>
<td>8 way, 32KB/core, private</td>
</tr>
<tr>
<td>L2</td>
<td>8 way, 256KB/core, private</td>
</tr>
<tr>
<td>L3</td>
<td>20 way, 20MB, shared</td>
</tr>
<tr>
<td>Memory</td>
<td>DDR 4, 1867MHz, SMI 2 channels</td>
</tr>
</tbody>
</table>

The negative impact of the GC latency increases with the number of compute nodes. As reported in [35], a GC run on a single node can hold up the entire cluster — when a node requests a data partition from another server that is running
Panthera: Holistic Memory Management for Big Data ...

...GC, the requesting node cannot do anything until the GC is done on the second node. Since Panthera can significantly improve the GC performance on NVM, we expect Panthera to provide even greater benefit when Spark is executed on a large NVM cluster.

**System Configurations.** Each CPU has 128GB DRAM. We reserved at least 8GB of DRAM for the OS and, hence, the maximum amount of DRAM that can be used for Spark is 120GB. We experimented with two different heap sizes for the Spark-running JVM (64GB and 120GB) and three different DRAM sizes (1/4, 1/3, and 100% of the heap size). The configuration with 100% DRAM was used as a baseline for us to compute the overhead of Panthera under hybrid memories.

Prior works on NVM often used smaller DRAM ratios in their configurations. For example, Write Rationing [9] used 1GB DRAM and 32GB NVM in their experiments. However, as we deal with Big Data systems, it would not be possible for us to use a very small DRAM ratio — in our experiments, a regular RDD consumes 10-30GB memory, and hence, we had to make DRAM large enough to hold at least one RDD.

The nursery space is placed entirely in DRAM. We have experimented with several different sizes (1/4, 1/5, 1/6, and 1/7 of the heap size) for the nursery space. The performance differences between the 1/4, 1/5, and 1/6 configurations were marginal (even under the original JVM), while the configuration of 1/7 led to worse performance. We ended up using 1/6 in our experiments to achieve good nursery performance and simultaneously leave more DRAM to the old generation.

**Programs and Datasets.** We selected a diverse set of 7 programs. Table 4 lists these programs and the datasets used to run them. These are representative programs for a wide variety of tasks including data mining, machine learning, graph and text analytics. PR, KM, LR, and TC run directly on Spark; CC and SSSP are graph programs running on GraphX [22], which is a distributed graph engine built over Spark; BC is a program in MLlib, a machine learning library built on top of Spark. We used real-world datasets to run all the seven programs. Note that although the sizes of these input datasets are not very large, there can be large amounts of intermediate data generated during the computation.

**Table 4. Spark programs and datasets.**

<table>
<thead>
<tr>
<th>Program</th>
<th>Dataset</th>
<th>Initial Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphX-Connected Components (CC)</td>
<td>Wikipedia Full Dump, English [3]</td>
<td>5.7GB</td>
</tr>
<tr>
<td>MLlib-Naive Bayes Classifiers (BC)</td>
<td>KDD 2012 [1]</td>
<td>10.1GB</td>
</tr>
</tbody>
</table>

**Baselines.** Our initial goal was to compare Panthera with both Espresso [54] and Write Rationing [9]. However, neither of them is publicly available. Espresso proposes a programming model for developers to develop new applications. Applying it to Big Data systems would mean that we need to rewrite each allocation site, which is clearly not practical. In addition, Espresso does not migrate objects based on their access patterns.

The Write Rationing GC has two implementations: Kingsguard-Nursery (KN) and Kingsguard-Writes (NW). KN places the young generation in DRAM and the old generation in NVM. KW also places the young generation in DRAM. Different from KN, KW monitors object writes and dynamically migrates write-intensive objects into DRAM. Although we could not directly compare Panthera with these two GCs, we have implemented similar algorithms in OpenJDK. Under KW, almost all persisted RDDs were quickly moved to NVM. The frequent NVM reads from these RDDs, together with write barriers used to monitor object writes, incurred an average of 41% performance overhead for our benchmarks. This is because Big Data applications exhibit different characteristics from regular, non-data-intensive Java applications.

KN appears to be a good baseline at the first sight. However, implementing it naïvely in the Parallel Scavenge collector can lead to non-trivial overhead — the reduced bandwidth in NVM can create a huge impact on the performance of a multi-threaded program; this is especially the case for Parallel Scavenge that attempts to fully utilize the CPU resources to perform parallel object scanning and compaction.

To obtain a better baseline, we placed the young generation in DRAM and support the old generation with a mix of DRAM and NVM. In particular, we divided the virtual address space of the old generation into a number of chunks, each with 1GB, and used a probability to determine whether a chunk should be mapped to DRAM or NVM. The probability is derived from the DRAM ratio in the system. For example, in a system with DRAM to memory ratio is 1/4 (1/4 DRAM), each chunk is mapped to DRAM with a 1/4 probability to NVM with a 3/4 probability. Note that this is a common practice [21, 51] to utilize the combined bandwidth of DRAM and NVM. We refer to this configuration as *unmanaged* in the rest of the section, which outperforms both KN and KW for our benchmarks.

**5.3 Performance and Energy.**

Figure 4 reports the overall performance and energy results when a 64GB heap is used and DRAM to memory ratio is 1/3 (1/3 DRAM). The performance and energy results of each configuration are normalized w.r.t. those of the 64GB DRAM-only version. Compared to the DRAM-only version, the unmanaged version reduces energy by 31.0% with a 21.4% execution time overhead. In contrast, Panthera reduces energy by 37.4% at a 4.3% execution time overhead.

When the heap size is 120 GB (not shown in Figure 4, but summarized later in Figure 6 and Figure 7), the unmanaged version reduces energy by 43.5% at a 19.3% execution time overhead. In contrast, Panthera reduces energy by 51.7% with less than 1% execution time overhead. Clearly, considering...
the RDD semantics in data placement provides significant benefit in both energy and performance.

**GC Performance.** To understand the GC performance, we broke down the running time of each program into the mutator and GC time; these results (under the 64GB heap) are shown in Figure 5. Compared to the baseline, the unmanaged version introduces performance overhead of 60.4% and 6.9% in the GC and computation, respectively; while for Panthera these two overheads are, respectively, 4.7% and 4.5%. Under the 120GB heap, the GC performance overhead of the unmanaged version and Panthera are, respectively, 58.0% and 3.1%. Note that due to large amounts of intermediate data generated, the GC is frequently triggered for these programs.

![Figure 4. Overall performance and energy results under a 64GB heap; DRAM to memory ratio is 1/3.](image)

Since the GC is a memory-intensive workload, inappropriate data placement can lead to significantly increased memory access time and thus a large penalty. The penalty comes from two major sources. First, NVM’s limited bandwidth (which is about 1/3 of that of DRAM) has a large negative impact on the performance of Parallel Scavenge, which launches 16 threads to perform parallel tracing and object copying in each (nursery and full-heap) GC. Given this high degree of parallelism, the performance of the nursery GC is degraded significantly when scanning objects in NVM. Second, object tracing is a read-intensive task, which suffers badly from NVM’s higher read latency.

Panthera improves the GC performance by pretenuring frequently-accessed RDD objects in DRAM and performing optimizations including *eager promotion* ($\S 4.2.2$) and *card padding* ($\S 4.2.3$). *Eager promotion* reduces the cost of (old-to-young) tracing in each minor GC, while *card padding* eliminates unnecessary array scans in NVM, which are sensitive to both latency and bandwidth. A further breakdown shows that *eager promotion*, alone, contributes an average of 9% of the total GC performance improvement. The contribution of *card padding* is much more significant — without this optimization, the GC time increases by 60% due to the impact of NVM’s substantially limited bandwidth and increased latency on the performance of parallel card scanning. In fact, this impact is so large that the other optimizations would not work well when card padding is disabled.

**Varying Heaps and Ratios.** To understand the impact of the heap sizes and DRAM ratios (DRAM to total memory), we have conducted experiments with two heap sizes (64GB, 120GB) and two DRAM ratios (1/3, 1/4) on four programs PR, LR, CC, and BC. Figure 6 reports the time results of these configurations. Panthera’s time overheads are, on average, 9.5%, 3.4%, 2.1%, and 0%, respectively, under the four configurations (64GB, 1/4), (64GB, 1/3), (120GB, 1/4), and (120GB, 1/3). The overheads for the unmanaged version are 25.9%, 20.9%, 23.9%, and 19.3%, respectively, under these same four configurations.

![Figure 5. Comparison of the GC performance (64GB heap).](image)

![Figure 6. Normalized times for two DRAM ratios + two heaps.](image)

We make two interesting observations. First, Panthera is more sensitive to the DRAM ratio than the heap size. The time overhead can be reduced by almost 10% when the DRAM ratio increases from 1/4 to 1/3. The reason is that more frequently accessed RDDs are moved to DRAM, reducing the memory latency and bandwidth bound of NVM. Another observation is that the unmanaged version is much less sensitive to DRAM ratio — the time overhead is reduced by only 5% when the DRAM ratio increases to 1/3. This is because arbitrary data placement leaves much frequently-accessed data in NVM, making CPUs stall heavily when accessing NVM.

Figure 7 depicts the energy results for the two heaps and two DRAM/NVM ratios. For the 64GB heap, the unmanaged version reduces energy by an average of 36.7% and 30.7%, respectively, under the 1/4 and 1/3 DRAM ratio, while Panthera reduces energy by 41.7% and 38.0% under these same
Panthera: Holistic Memory Management for Big Data ...

ratios. The energy reductions for the 120GB heap are much more significant — the unmanaged version reduces energy by 50.2% and 43.5%, respectively, under the 1/4 and 1/3 DRAM ratios, while the energy reduction under Panthera increases to 57% and 51.7% for these two ratios.

![Figure 7. Normalized energy for two DRAM ratios + two heaps.](image)

**Figure 7.** Normalized energy for two DRAM ratios + two heaps.

### 5.4 Memory Access Analysis

NVM has high latency and low bandwidth. In general, the performance penalty caused by high latency increases with the number of memory access. For the same number of memory access, NVM incurs more performance penalty for applications which has lots of high instant bandwidth usage beyond the limits of NVM bandwidth.

Figure 8 depicts the read/write bandwidth between unmanaged, and Panthera for GraphX-CC. Compared to unmanaged version, Panthera migrates most of the memory read/write from NVM to DRAM and eliminates lots of high instant memory access bandwidth (peaks in figure). Because Panthera allocates/moves frequently accessed data to DRAM and reduces lots of unnecessary memory access (§4.2.2, §4.2.3).

![Figure 8. GraphX-CC’s memory access bandwidth.](image)

**Figure 8.** GraphX-CC’s memory access bandwidth.

### 5.5 Overhead of Monitoring and Migration

As discussed in §4.2, Panthera performs lightweight method-level monitoring on RDD objects to detect mis-placed RDDs for dynamic migration. This subsection provides a closer examination of dynamic migration’s overhead.

As we monitor only method calls invoked on RDD objects, we find its monitoring overhead is negligible: it is less than 1% across our benchmarks. For example, for PageRank, only about 300 calls were observed on all RDD objects in a 20-minute execution. The second column of Table 5 reports the number of calls monitored for each application. For GraphX applications, which has thousands of RDD calls, the monitoring overheads are still under 1%.

Dynamic migration (performed by the major GC) rarely occurs in our experiments, as can be seen from the third column of Table 5. There are two main reasons. First, the frequency of a major collection is very low because a majority of objects die young and most of the collection work is done by the minor GC. Second, for four applications (PR, KM, TC, and LR), our static analysis results are accurate enough and, hence, dynamic migration is never needed.

We observed that only two RDDs (during the executions of CC and SSSP) were migrated dynamically. Note that both CC and SSSP are GraphX applications. Each iteration of the processing creates new RDDs representing the updated graph and persists them. At the end of each iteration, the RDDs representing the old graph are explicitly unpersisted. Our static analysis, due to lack of support for unpersisted, marks both old and new graph RDDs as hot data and generates a DRAM tag for all them. These RDD objects are then allocated in DRAM and their data objects are promoted eagerly to the DRAM space of the old generation. The RDD objects representing the old graphs, if they can survive a major GC, are migrated to the NVM space of the old generation due to their low access frequency.

To have better understanding of the individual contributions of pretenuring and dynamic migration, we have disabled the monitoring and migration and rerun the entire experiments. The performance difference was not noticeable. Hence, we conclude that most of Panthera’s benefit stems from pretenuring, which improves the performance.

---

**Table 5.** Dynamic monitoring and migration.

<table>
<thead>
<tr>
<th>Program</th>
<th># Calls monitored</th>
<th># RDDs migrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR</td>
<td>328</td>
<td>0</td>
</tr>
<tr>
<td>KM</td>
<td>550</td>
<td>0</td>
</tr>
<tr>
<td>LR</td>
<td>333</td>
<td>0</td>
</tr>
<tr>
<td>TC</td>
<td>217</td>
<td>0</td>
</tr>
<tr>
<td>CC</td>
<td>2945</td>
<td>1</td>
</tr>
<tr>
<td>SSSP</td>
<td>3632</td>
<td>1</td>
</tr>
<tr>
<td>BC</td>
<td>336</td>
<td>0</td>
</tr>
</tbody>
</table>
of both the mutator and the GC. However, dynamic monitoring and migration increases the generality of Panthera’s optimizations, making Panthera applicable to applications with diverse access characteristics.

6 Related Work

Hybrid Memories for Managed Runtime. To our knowledge, Panthera is the first practical work to optimize data layout for managed-runtime-based distributed Big Data platforms. Existing efforts [9, 12, 21, 25, 26, 43, 48, 50, 54] that attempt to support persistent Java focus on regular applications or need to totally rebuild the platforms.

Inoue and Nakatani [24] identify code patterns in Java applications that can cause cache misses in L1 and L2. Gao et al. [20] propose a framework including support from hardware, the OS, and the runtime to extend NVM’s lifetime. Two recent works close to Panthera are Espresso [54] and Write Rationing [9]. However, they were not designed for Big Data systems. Espresso is a JVM-based runtime system that enables persistent heaps. Developers can allocate objects in a persistent heap using a new instruction _pnew_ while the runtime system provides crash consistency for the heap. Applying Espresso needs to rewrite the Big Data platforms (e.g., Spark) by _pnew_, which isn’t practical.

Write Rationing [9] is a GC technique that places highly mutated objects in DRAM and read-only objects in NVM to increase NVM lifetime. Like Espresso, this GC focuses on individual objects and does not consider application semantics. Panthera’s nursery space is also placed in DRAM, similar to the Kingsguard-nursery in Write Rationing. However, instead of focusing on individual objects, Panthera utilizes Spark semantics to obtain access information at the array granularity, leading to effective pretenuring and efficient runtime object tracking.

Memory Structure. There are two kinds of hybrid memory structures: _flat structure_, where DRAM and NVM share a single memory space, and _vertical structure_, where DRAM is used as a buffer for NVM to store hot data. The vertical structure is normally managed by hardware and transparent to the OS and applications [27, 31, 34, 36, 45, 57, 60]. Qureshi et al. [45] shows that a vertical structure with only 3% DRAM can reach similar performance to its DRAM-only version. However, the overhead of page monitoring and migration increases linearly with the working set [51]. The space overhead _e.g._, the tag store space of DRAM buffer, can also be high with a large volume of NVM.

Page-based Migration. A great number of existing works use memory controllers to monitor page read/write frequency [15, 16, 19, 23, 33, 44, 46, 51, 59] and migrate the top-ranked pages to DRAM. Another type of hybrid memory, composed of 3D-stacked DRAM and commodity DRAM, also adapts similar page monitoring policies [17]. However, none of these techniques were designed for Big Data systems. Hassan et al. [23] show that, for some applications, migrating data at the object level can reduce power consumption.

For Big Data applications that have very large memory consumption, continuous monitoring at the page granularity can incur an unreasonable overhead. Page migration also incurs overhead in time and bandwidth. Bock et al. [13] report that page migration can increase execution time by 25% on average. Panthera uses static analysis to track memory usage at the RDD granularity, incorporating program semantics to reduce the monitoring overheads.

Static Data Placement. There exists a body of work that attempts to place data directly in appropriate spaces based either on their access frequencies [15, 33, 44] or on the result of a program analysis [19, 23, 51]. Access frequency is normally calculated using a static data liveness analysis or offline profiling. Chatterjee et al. [15] place a single cache-line across multiple memory channels. Critical words (normally the first word) in a cache-line are placed in a low-latency channel. Wei et al. [51] show that the group of objects allocated by the same site in the source code exhibit similar lifetime behavior, which can be leveraged for static data placement.

Li et al. [33] develop a binary instrumentation tool to statistically report memory access patterns in stack, heap, and global data. Phadke and Narayanasamy [44] profile an application’s MLP and LLC misses to determine from which type of memory the application could benefit the most. Kim et al. [28] develop a key-value store for high-performance computers with massive NVM, which provides developers with a high-level interface to use the distributed NVM. However, none of these techniques were designed for managed Big Data systems.

7 Conclusion

We present Panthera, the first memory management technique for Big Data processing over hybrid memories. Panthera combines static analysis and GC techniques to perform semantics-aware data placement in hybrid memory systems. Our evaluation shows that Panthera reduces energy significantly without incurring much extra time overhead.

Acknowledgments

We thank the anonymous reviewers for their thorough and insightful comments. We are especially grateful to our shepherd Jennifer Sartor for her feedback, helping us improve the paper substantially. The work is supported in part by the National Key R&D Program of China grant 2016YFB1000201, National Science Foundation grants CNS-1613023, CNS-1703598, and CNS-1763172, Office of Naval Research grants N00014-16-1-2913 and N00014-18-1-2037, National Natural Science Foundation of China grants 61802368, 61521092, 61432016, 61432018, 61332009, 61702485, and 61872043, as well as gifts from Alibaba, Intel and VMware.
References


