

Arcadia: A Fast and Reliable Persistent Memory Replicated Log

Shashank Gugnani^{*}
The Ohio State University

Scott Guthridge
IBM Research

Frank Schmuck
IBM Research

Owen Anderson
IBM Research

Deepavali Bhagwat[†]
IBM Research

Xiaoyi Lu
UC Merced

Abstract

The latency and bandwidth properties of byte-addressable persistent memory (PMEM) have the potential to significantly improve system performance over a wide spectrum of applications. But at the same time, PMEM brings considerable new challenges for the programmer compared with earlier technologies. In particular, PMEM provides only 8-byte write atomicity, can flush writes out of order, and its availability is always limited by node failures. It's possible to work with the atomicity and ordering constraints of PMEM directly by carefully sequencing the order of store operations and inserting explicit flush and fence operations at each ordering point. But this is tedious and error-prone: too many flush operations defeat the performance benefits of PMEM, and even with generous use, it is difficult to prove that a given program is crash-consistent. Logging is a great abstraction to deal with these issues but prior work on PMEM logging has not successfully hidden the idiosyncrasies of PMEM. Moreover, shortcomings in the log interface and design have prevented attainment of full PMEM performance. We believe that a log design that hides the idiosyncrasies from programmers while delivering full performance is key to success. In this paper, we present the design and implementation of Arcadia, a generic replicated log on PMEM to address these problems. Arcadia handles atomicity, integrity, and replication of log records to reduce programmer burden. Our design has several novel aspects including concurrent log writes with in-order commit, atomicity and integrity primitives for local and remote PMEM writes, and a frequency-based log force policy for providing low overhead persistence with guaranteed bounded loss of uncommitted records. Our evaluation shows that Arcadia outperforms state-of-the-art PMEM logs, such as PMDK's libpmemlog, FLEX, and Query Fresh by several times while providing stronger log record durability guarantees. We expect Arcadia to become the leading off-the-shelf PMEM log design.

^{*}Now at Oracle

[†]Now at Google

1 Introduction

Persistent Memory (PMEM) provides byte-addressable access to data and persistence across power cycles with latency and bandwidth properties that make it an attractive storage technology. At nearly the speed of DRAM, it promises significant performance gains for a number of important workloads. Unfortunately, the persistence of PMEM is hard to realize in practice [19, 31]. Until store operations have been flushed from CPU caches, they remain volatile, thus are lost on power failure. Moreover, the CPU can reorder stores and implicitly evict cache lines at any time, making it necessary for the programmer to meticulously order stores and explicitly flush cache lines in order to maintain crash consistency.

Even then, the hardware provides only 8-byte atomicity for local writes and so far *no* atomicity for remote writes over the network, further complicating the work of updating data in a crash-consistent way [15]. Using the memory interface for speed also makes PMEM intrinsically *local*, thus node failures and network partitions curtail its availability. This is in stark contrast with the 4KB sector atomicity typical of other common block-based storage technologies and the ability to share storage across a storage area network such as NVMe-oF.

A further complication not unique to PMEM is that uncorrectable media errors or DIMM failures may result in permanent loss of data [47].

All of these problems can be solved by introducing a replicated log. First, logging solves the problem of atomicity of updates for crash consistency across power failures. Before making changes to PMEM, the programmer first records into the log a description of the changes to be made (typically redo or undo records), then makes the changes to primary storage in PMEM. If the system crashes while making the changes to PMEM, log recovery at the next start-up restores consistency of the data in PMEM by completing (with redo) or reverting (with undo) any partial updates [30]. Second, replicating the log across nodes solves the problems of node failures, network partitions, uncorrectable media errors, and PMEM DIMM failures.

To preserve the performance benefits of PMEM as fast primary storage, the log itself must also reside in fast PMEM storage. The log must be implemented with performance and reliability in mind and must be able to handle all the programming challenges of PMEM. To meet these, a lot of work has to go into the design of the log. But once this has been done, the log provides a platform on which the programmer can easily realize the benefits of PMEM as primary storage with full performance and reliability.

In addition to serving as a log for PMEM as primary storage, the fast PMEM log can also provide crash consistency and excellent update latency for primary storage on any other persistent medium, reducing the log latency that often limits update rate.

PMEM’s byte addressability is an excellent match with remote direct memory access (RDMA). With RDMA operations, data can be transferred quickly to and from PMEM on a remote node. Using RDMA with PMEM is not without complications, however, because it suffers from many of the same atomicity and persistence problems as local PMEM. Thus, the abstractions the log uses to manage the challenges of PMEM locally must be extended to PMEM over RDMA.

Why doesn’t prior work solve all problems? Prior work on PMEM logging [4, 13, 17, 22, 36, 42, 43, 45] places inordinate trust in hardware to retain and atomically update data. Prior works were designed with the assumption that PMEM has reliability similar to DRAM. However, reliability for PMEM is a more complex and underappreciated problem¹: media errors and software bugs can *permanently* corrupt data, and the reliability of each memory-cell degrades as it is used, potentially leading to premature failure [34, 47]. So, in practice, without sufficient redundancy and integrity checking, data may be lost or corrupted, putting additional burden on the programmer to achieve reliability. Further, prior work does not adequately provide write concurrency necessary for full PMEM performance while also ensuring in-order commit (monotonicity).

In this paper, we present the design and implementation of **Arcadia**², a generic replicated log to tame the challenges of PMEM. We make the following novel contributions in this paper:

1. A generic interface for Arcadia that has two distinguishing features: it hides the complexities of replication and persistence, and it decouples steps that must be serialized from those allowing concurrency. Log append concurrency provides for application processing overlap, hides the latencies of checksum calculation and replication, and provides concurrency necessary for full performance.
2. Persistence, atomicity, and integrity primitives for both

local and remote accesses to PMEM. These abstractions hide the weak consistency of PMEM, ensure reliable persistence, and make PMEM easier to use.

3. A design and implementation using the primitives above to handle atomicity, integrity, and replication of log records with total-order semantics.
4. A frequency-based log force policy for providing low overhead persistence with guarantees on bounded data loss.

We deploy Arcadia on two servers with Intel Optane PMEM that are connected with RDMA. Our analysis shows that Arcadia significantly outperforms state-of-the-art PMEM logs, such as FLEX [43], PMDK’s libpmemlog [17], and Query Fresh [42] while providing stronger log record durability guarantees. We expect Arcadia to be used as an off-the-shelf log implementation for any storage system using logging, in particular systems that have weak consistency or disaggregated storage.

2 Background and Motivation

In this section, we discuss the background and motivation of our work.

2.1 Persistent Memory

PMEM technologies, such as PCM [23, 24], ReRAM [2], STT-RAM [39], 3D-XPoint [10], and memristor [35] are disruptive technologies in storage system design. With three orders of magnitude better latency and an order of magnitude better bandwidth compared to flash [29, 46], PMEM is an excellent candidate for storing latency critical data. Given that PMEM is an emerging technology with much higher cost than flash [21], we do not expect PMEM to replace flash as primary storage in the foreseeable future. Rather, we expect it to be used in the role of latency critical logging and metadata storage in next generation systems.

Despite its performance benefits, the write atomicity and ordering problems of PMEM make it difficult to achieve both performance and crash consistency.

2.2 RDMA Networking

Modern networking interconnects such as InfiniBand [14] and RDMA over Converged Ethernet (RoCE) [5] are heavily used in high-performance computing to achieve high throughput and low latency. These technologies are now being increasingly adopted in data centers worldwide. Through RDMA, a process can read or update memory of a remote process while minimizing remote involvement. Data transmission bypasses both the local and remote OS, allowing zero-copy operation with data transferred directly between network and memory controllers. Protocol processing can generally be offloaded

¹With DRAM, if there is data corruption or a media error, simply restarting the application solves the problem, but this is not possible with PMEM.

²Named after the Greek province of Arcadia.

to the hardware, further improving performance. RDMA provides both two-sided (Send and Recv) and one-sided (Read and Write) primitives. The one-sided primitives benefit from being able to transfer data without any critical path involvement of the remote side. Choosing the appropriate primitive is necessary for high-performance and low latency communication.

2.3 PMEM Persistence over RDMA

As with DRAM, RDMA can be used to directly access data in PMEM. Unfortunately, the one-sided RDMA primitives were not designed for the complicated persistence properties of PMEM. RDMA writes to PMEM are acknowledged as soon as data has reached the remote NIC. So, there is no guarantee that data has been written to the desired memory location as it could be cached within the NIC or PCIe buffers. Once data has been transferred out of the NIC, there is still no guarantee that it has been made persistent because modern NICs can put data directly into volatile CPU caches, still outside of the PMEM persistence domain. As a result of the complicated write path, there is no persistence or atomicity guarantees for RDMA writes to PMEM. This makes the design of a replicated log challenging because a fundamental requirement of logging is to make record updates atomic and persistent. One solution for solving both atomicity and persistence issues is to leverage two-sided RDMA primitives in a request and response replication model with the remote side responsible for persisting data before responding to the replication request [8]. In doing so, the performance benefits of one-sided RDMA cannot be realized. Another solution is to leverage special hardware primitives, such as RDMA Commit [37] or to always force the NIC to write data to PMEM instead of CPU caches. The former relies on hardware support from the NIC that is currently not available in any commodity product, while the latter requires BIOS configuration changes that also penalize the performance of any network traffic not destined for PMEM. Therefore, one-sided RDMA writes cannot directly be leveraged to replicate data to remote PMEM unless both the persistence and atomicity limitations are addressed.

2.4 Limitations of Prior Work

The performance critical nature of logging has led many researchers to propose new log implementations on PMEM as a way of improving end-to-end system performance. Table 1 presents a comparison of resilience to key failure scenarios between Arcadia and related work. Several works have looked at designing both unreplicated and replicated logging protocols on PMEM [4, 13, 17, 22, 36, 42, 43, 45, 48]. Unreplicated protocols, like PMDK’s libpmemlog [17] and FLEX [43] are by design not resilient to node or device failures. This makes it hard to apply them in systems with reliable and replicated primary storage. The log is a single point of failure in persistent systems, so its resilience must be at least that of primary storage; therefore, replicated logging is essential for a robust

Log Design	Device/Node Failure	Network Partition	Media Error	Power Loss
PMDK [17]	✗	✗	✗	✓
FLEX [43]	✗	✗	✗	✓
Query Fresh [42]	✓	✓	✗	✓
Tailwind [36]	✓	✓	✓	✗
Arcadia	✓	✓	✓	✓

Table 1: Comparison of resilience to key failure scenarios with related work. Note that Arcadia is the only system that is robust to all of these failure scenarios. In addition, it is the only one to provide concurrent writes with in-order commit.

storage system. Unfortunately, prior work on replicated logging places more trust in PMEM hardware’s ability to retain data than is prudent. This is because these works were designed with the assumption that PMEM has similar reliability to DRAM. However, reliability for PMEM is more complex: media errors and software bugs can permanently corrupt data, and the reliability of each memory-cell degrades as it is used, potentially leading to premature failure [34, 47]. So, in practice, without sufficient redundancy and integrity checking, data may be lost or corrupted, pushing the burden of reliability onto the programmer. For instance, Query Fresh [42] and Mojim [48] do not have any mechanism to handle memory corruption (resulting from software accidentally writing data to the wrong place) or undetected media errors, so it is possible for clients to read silently corrupted data. Further, Mojim does not offer concurrent log writes. Tailwind [36] assumes the presence of DMA capable battery-backed DRAM buffers to guarantee persistence. Such special hardware support is impractical in a production system and we are not aware of any commodity system offering such support.

Providing concurrency for log writes on PMEM is another dimension where prior work falls short. It is challenging to provide concurrent writes to the log because multi-threaded concurrency is non-deterministic and storage systems typically require log writes to preserve a total order (monotonicity) to guarantee correctness. Without a total order, consistency of systems cannot be guaranteed on recovery. While some prior work does provide concurrent log writes, those designs are for block-based disk or flash storage, not for byte-addressable PMEM. PMEM makes this problem more challenging because cache lines may be evicted implicitly at any time. So, if log data is accessed concurrently and in-place, there are no ordering guarantees for data persistence. Prior work completely sidesteps this issue by fully isolating log writers to preserve the total log order but at the cost of limited or no concurrency.

Arcadia was designed to overcome these shortcomings of prior work. Through replication, Arcadia can survive device/node failures and network partitions. Further, by leveraging novel integrity and atomicity primitives for PMEM, Arcadia is resilient to memory corruption, media errors, and power loss.

Arcadia also provides concurrency while preserving in-order commits. To achieve this goal, Arcadia separates log writes into a set of distinct steps and *only* isolates those that require serialization. In this manner, unnecessary synchronization can be avoided and concurrency is maximized.

3 Primitives for Local and Remote Accesses

Handling atomicity, integrity, and persistence of accesses to PMEM is a non-trivial task. In this section, we first present primitives for durable local and remote writes to PMEM. Building upon these, we present primitives for atomically and reliably accessing data in local or remote PMEM. Crucially, these primitives handle the complexities of PMEM + RDMA persistence and hide the weak and complicated persistence properties of PMEM. We distinguish the primitives as Persistence, Replication, Integrity and Atomicity.

Use Cases. The proposed primitives help make PMEM easier to use. The integrity primitive can be used to reliably write data once, or when the data being overwritten is known not to be needed anymore. The atomicity primitive can be used when updating existing data in place, where we need to make sure that if there is a failure during write, the old data remains readable. We show how these primitives can be applied in Arcadia to achieve the persistence, integrity, and atomicity requirements of log writes. The persistence and replication primitives are building blocks for the integrity and atomicity primitives, which Arcadia uses to format different portions of the log – the integrity primitive for log records and the atomicity primitive for the log header. Additional details on how Arcadia uses these primitives can be found in §4.3.

Persistence Primitive. To guarantee persistence locally, we construct a primitive using hardware access macros. It takes a location and length in PMEM and makes sure that data previously stored at that location is persistent. It relies on architecture-specific instructions such as `clwb` and `sfence` that can guarantee that data has been flushed from the volatile CPU caches into PMEM.

Replication Primitive. Our goal is to construct a replication primitive which can work on currently available commodity hardware and does not require hardware-modification or BIOS configuration changes. To this end, we design a single round-trip protocol for both replicating and persisting data to remote PMEM. We use the hybrid RDMA-Write-with-Immediate (RDMA-Write-Imm) operation to both replicate data and signal the remote server to force data to storage using the persistence primitive. The ‘immediate’ value in this operation carries with it the length of the data to force while its starting address is obtained from the RDMA-Write completion event. Essentially, the RDMA-Write-Imm acts as an asynchronous RPC to signal the remote server to persist data. Once the persist operation is complete, the remote server sends an acknowledgement using a (two-sided) RDMA Send. The local server can use this acknowledgement as an

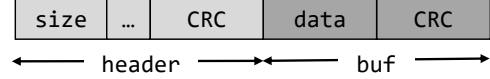


Figure 1: Integrity Primitive Data Layout in PMEM

Listing 1: Integrity Primitive

```

1 # header and buf are preallocated in PMEM
2 function ReliableWrite(data, size)
3     header = generateHeader(data, size)
4     header.crc = crc32(header) # Header CRC
5     memcpy(buf.data, data, size) # Copy data
6     buf.crc = crc32(data) # Data CRC
7     # Replicate + Force header and data
8     return rdma_write_and_force(header, buf)
9
10 function ReliableRead(data)
11     rdma_read(header, buf) # Remote read
12     if (header.crc != crc32(header))
13         return false
14     if (buf.crc != crc32(buf.data))
15         return false
16     memcpy(data, buf.data, header.size)
17     return true

```

assurance of remote data persistence. This approach preserves the asynchrony of one-sided RDMA while also guaranteeing remote persistence.

Integrity Primitive. This primitive is designed for reliably reading and writing data in PMEM (e.g., log records). Reliability here means that we should be able to verify the integrity of data. This requires protection against two situations – the first where writes may be torn, i.e., only part of data is persisted, and the second where data is corrupted because of undetected media errors. To achieve reliability, we add a header field to every data item with checksums protecting both of header and data. Figure 1 provides an overview of the data layout for this primitive. The header contains the size of the data buffer and can be used to store additional information, for instance, to identify the type of data stored. Listing 1 presents a pseudo-code of the primitive. The `rdma_write_and_force()` function represents the replication primitive and can be replaced by the persistence primitive if replication is not desired. For reliably writing data, checksums (such as CRC32) for header and data are generated. By protecting both using checksums, there are no ordering requirements of writes or persistence barriers for the header and data. In addition, there is no requirement on atomicity of writes to PMEM. This is crucial because replicating data using one-sided RDMA does not provide any atomicity guarantees. In our approach, a single replicate and persist operation is sufficient for the entire write. When reading data, both the header and data checksum must be checked before data can be safely copied. Header integrity must be validated before data integrity, otherwise the size field in the header may not be correct. As an optimization in some circumstances, the header checksum can be replaced by a special integrity check

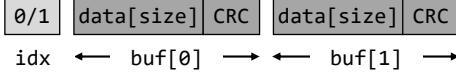


Figure 2: Atomicity Primitive Data Layout in PMEM

Listing 2: Atomicity Primitive

```

1  # idx and buf[] are preallocated in PMEM
2  function AtomicWrite(data)
3      memcpy(buf[!idx].data, data, size)
4      buf[!idx].crc = crc32(data) # Generate CRC
5      # Replicate + Force data and CRC
6      rdma_write_and_force(buf[!idx])
7      idx = !idx # Flip index
8      # Replicate + Force index
9      return rdma_write_and_force(idx)
10
11 function AtomicRead(data)
12     # Remote read
13     rdma_read(idx, buf)
14     if (buf[idx].crc != crc32(buf[idx].data))
15         return false
16     memcpy(data, buf[idx].data, size)
17     return true

```

value, such as a log sequence number (LSN), to eliminate the cost of generating the checksum.

Atomicity Primitive. This primitive is designed for atomically accessing data in PMEM and is required when updating an object with a fixed location (e.g., the log’s header). It guarantees that the entire data is updated atomically and that data integrity can be validated on reads. Atomically updating data is challenging because writes may be torn, so data may be only updated partially. Data cannot be updated in-place because there is no guarantee of atomicity for remote writes to PMEM. To solve this issue, we propose the use of copy-on-write (CoW) for updates to data. Figure 2 provides an overview of the data layout and Listing 2 presents a pseudo-code of the primitive. We use two buffers to implement the CoW approach and switch between the two for every update. The index flag identifies the current valid buffer for reads. Checksums protect data integrity. For any update, data must be written to the invalid buffer. Once data and its checksum have been written and persisted, the index can be updated and persisted accordingly. As an optimization, to eliminate the cost of persisting the index flag, the flag can be placed in volatile storage as long as the valid data buffer can be identified on recovery using the contents of the data. Another possible optimization, to reduce space usage, is to always write to a new dynamically allocated buffer and discard the old buffer once the write has successfully completed.

4 Design

In this section, we present the design and implementation of Arcadia. We design Arcadia as a replicated log with a single-primary, multi-backup model. It has a single multi-

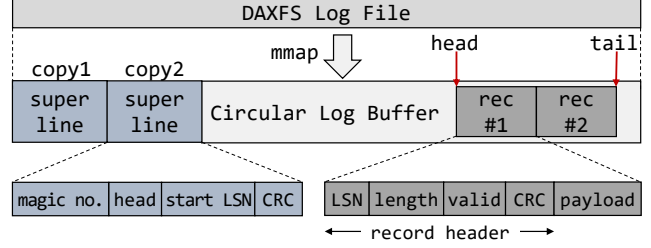


Figure 3: Arcadia’s Layout on PMEM

threaded writer doing updates (logger) and a single reader during recovery.

4.1 Design Overview

PMEM Layout. We place the entire log as a single file on a DAX filesystem formatted on PMEM. The entire file is `mmap`ed into the address space of the process to enable load/store access to log records. Figure 3 shows an overview of the log layout on PMEM. The log itself consists of a circular buffer of fixed size to hold the log records and a header containing information necessary to both identify and recover instances of the log. We call this header the *superline*, a term combining superblock with cache line. The superline stores the head and start LSN of the log, which recovery uses to find the first valid record. We do not include the log tail in the superline to avoid the overhead of having to update it on each log append. Instead, we find the log tail on recovery by iterating over all valid records to locate the end of the log. Each log record in Arcadia consists of a header and payload pair. The header contains the length and checksum of the payload as well as a monotonically increasing LSN and a valid flag. Arcadia relies on special values and basic consistency checks like magic numbers, LSNs, checksums, and valid flags to validate the log’s content in persistent storage.

Operation Modes. Arcadia can be deployed in one of three operation modes – *local*, *local+remote*, or *remote only*. In the *local* mode, log data is placed and accessed only on PMEM available on the local server. In *local+remote* mode, the local copy is the primary replica of the log with one or more secondary servers serving as backups. Finally, in *remote only* mode, all durable copies of the log reside on one or more remote backups, while the client only holds a volatile copy of the log.

Whenever Arcadia is configured in a mode with replication, at most one replica is local and the rest must be remote. Because remote replicas have higher latency than local due to network overhead, it follows that when replication is in use, write latency is always limited by the network, thus having a local replica does not significantly improve performance. A positive consequence of this is that log clients can be located anywhere there is good RDMA connectivity, and are not required to have local PMEM.

Interface. Arcadia’s interface was designed with three goals – (1) minimizing unintended synchronization, (2) providing

API Function	Description
<i>[id, ptr] reserve(size)</i>	Reserve space for a record
<i>int copy(id, data, size)</i>	Copy data into a record
<i>int complete(id)</i>	Mark a record as completed
<i>int force(id, freq=1)</i>	Force a record to PMEM
<i>id append(data, size)</i>	Append a record
<i>LSN getLSN(id)</i>	Get record LSN
<i>int cleanup(id)</i>	Reclaim record space
<i>int cleanupAll()</i>	Reclaim entire log
<i>ITERATOR begin/end()</i>	Recovery iterator

Table 2: Arcadia’s Interface

direct access to PMEM, and (3) enabling use in any situation requiring logging. An overview of the interface is presented in Table 2.

Conventionally, log interfaces only include an *append* method for log writes. Bundling the entire write as a single append method limits overlap of application compute with log writes and data persistence. To achieve the first goal, we propose a set of fine-grained methods to write data in the log. These methods break the append method into four separate stages – reserve, copy, complete, and force. This separation provides users flexibility to place the calls at the best possible location to maximize concurrency of application and log-related activity. A single append method is also provided as a combination of the four fine-grained methods wrapped up into a single convenient operation when the fine-grained interface isn’t needed.

The second goal is achieved by providing a direct pointer to the PMEM region allocated for a record using *reserve*. The benefit of having a direct pointer is that it allows the user to assemble the log record contents directly in PMEM without need to first build it in DRAM and then copy it. It avoids the need for a data copy in cases where the data to be logged is not already sitting in DRAM.

Finally, to achieve the third goal, we provide several generic methods to add records, reclaim space used by records, and iterate over valid records for recovery. These methods are described in detail in §4.3.

4.2 Replication and Recovery Protocol

Arcadia implements a quorum-based protocol for replication and recovery. The write quorum (W) is a configurable parameter that can be adjusted based on desired performance and fault-tolerance. The read quorum (R) is automatically selected based on W and the number of durable replicas (N) to satisfy quorum requirements ($R + W > N$). Therefore, the system can tolerate up to $N - W$ replica failures when writing log records.

Arcadia assumes that it is running within an existing cluster infrastructure (such as Apache Zookeeper [3]) that manages

membership and quorum of nodes, and that assigns an active *primary* node to control each instance of the Arcadia log and run the user application.

Replication. When log data needs to be replicated, RDMA Writes are initiated to all backups (remote log replicas) in parallel. Next, we wait for the persistence acknowledgement from all backups. Once sufficient writes have completed to meet the write quorum, the data can be considered to be durably replicated. Network partitions and backup failures can disrupt this process. To handle such cases, Arcadia uses a timeout-based approach. If the backup write times out, we consider the backup to be failed and immediately close the connection with that server. This also avoids the problem of an inconsistent backup in cases with a transient network partition. As long as the number of failed backups does not impact the write quorum (i.e., W backups are in complete sync), the replication process is successful. In situations where write quorum cannot be achieved, replication will fail, and as a consequence any call to *force* will also fail. We delegate responsibility for these failures to the application. An easy fix that the application can use in such cases is to gracefully shutdown, add new backup servers (by copying the PMEM log files), and restart the application.

Recovery. Once the existing cluster infrastructure has assigned an active node to manage the Arcadia log, Arcadia checks that a sufficient number of copies of the log are available and consistent to meet the log read quorum, R . If the read quorum is not met because backups are unavailable, Arcadia fails recovery and lets the user retry after more backups come online. If read quorum fails because too many copies are corrupted, then Arcadia cannot guarantee data consistency and recovery cannot proceed safely. During the recovery process, we use data in the superline to detect which copies of the log are consistent and contain the most recent data. Before accepting new requests, Arcadia repairs any inconsistent or failed log copies using the consistent copies. Only inconsistent copies are modified during recovery, therefore, the recovery process is idempotent and invulnerable to repeated failures. Once recovery has completed, the primary and all backups are up to date, and new requests can be safely accepted.

Handling Primary Failure. The primary node hosting the user application can fail at any time. We ensure that Arcadia handles such failures by using the cluster infrastructure. On primary failure, this infrastructure triggers leader election, selects a new primary, and informs all backups of the primary change. All backups immediately close their connection with the primary to ensure that it is fenced off. This prevents the old primary from continuing to replicate log records after it comes back up. The user application is then migrated to the new primary and restarted once log recovery is complete. Since Arcadia replicates log records synchronously, the new primary is guaranteed to have the latest log records that were forced by the client. This preserves both application consistency and correctness.

Handling Diverging Histories. Diverging histories can happen due to repeated failures of different backups. For example, assume we have 3 log replicas (*A*, *B*, and *C*) with $R=2$ and $W=2$. Replica *A* writes a record of value *X* at LSN 1, and then crashes. Recovery reads replicas *B* and *C*, which have not written anything at LSN 1. They are consistent and recovery proceeds without *A*. Next, replicas *B* and *C* write a record of value *Y* at LSN 1, and then crash. Recovery reads replicas *A* and *B* and discovers one replica with *X* at LSN 1, and one with *Y* at LSN 1, creating a conflict during recovery. To solve this problem, we add an epoch field to the superline in each log copy. The epoch values start out at 1. On each recovery, all available copies are read and the maximum epoch value is calculated. At least read quorum of copies must be readable, or recovery cannot continue. Recovery increments the epoch by 1 and writes the new value to all available copies. Writes to a write quorum of nodes must be successful to continue. Only log copies with the maximum epoch are considered valid. So, in the example above, only replicas *B* and *C* will be considered valid during the last recovery, thereby avoiding diverging histories.

4.3 Using the Log

In this sub-section, we describe how Arcadia accesses and updates the log in detail. Arcadia uses the proposed integrity primitive for accessing log records and the atomicity primitive for accessing the superline. As an optimization to the integrity primitive, we use the LSN for validating the header rather than a checksum. For optimizing the atomicity primitive, we keep the index in volatile memory and identify the valid superline on recovery based on which copy has the latest start LSN.

Concurrent Log Writes and Monotonicity. Arcadia allows concurrent writes despite having in-order commits. We describe how each write method is handled internally to enable this. The **reserve** method allocates space for log records and returns a direct pointer to the allocated space in PMEM³. It is also responsible for generating the LSN for the record. To ensure monotonicity of LSNs, this method synchronously allocates log space and generates LSNs. The **copy** method is a convenience method to copy data into a reserved record. It uses non-temporal stores that bypass the CPU caches to copy data to PMEM efficiently. This method can be safely called multiple times to copy different data chunks into the record. Note that users may choose not to use this method and instead use `memcpy` or CPU stores to copy data themselves. The **complete** method is used once all data has been written to the record. Complete generates and writes the checksum for the record payload and also sets the valid flag in the record header. The **force** method guarantees that a record is durable, i.e., it will be available after a crash. It does this by replicating and persisting records. Importantly, it ensures that there are no gaps in the stream of committed records. To do this, force

waits, if necessary, until all previous records have been marked valid (complete) and forces them first. Doing so ensures that records are always persisted in-order despite concurrent log writers.

The key insight here is that reserve and force are the only synchronous methods required during a log write. Multiple threads can safely call copy and complete in parallel. Although data is concurrently written into the log, the reserve and force methods carefully track the status of all records to ensure that monotonicity is still maintained.

Log Space Reclamation. Space reclamation is important when records are no longer necessary for recovery because the changes they describe have already been made durable. Depending on the type of logging algorithm used, records may be invalidated at different times. However, a cleanup method operating at the granularity of a record is sufficiently generic to be applicable in all cases. This method is responsible for unsetting the valid flag of the record. It also checks if there is a contiguous section of the log starting from the head (the oldest valid record) that can be reclaimed. If so, it advances the log head and updates its value in the superline. A `cleanupAll` method is also provided to reclaim the entire log. This method simply re-initializes the entire log and updates the superline accordingly, but preserves the current log epoch number.

Recovery Iterator. This iterator is used to access all valid records upon recovery from a crash. It is useful in bringing the system back to a consistent state using data in the log. Before it can be used, the recovery protocol (see §4.2) is triggered to repair any corruption and ensure consistency of all copies. The recovery iterator operates on the local log copy and obtains the log head from its superline to find the oldest valid record. To validate records, three integrity checks are performed – (1) all LSNs leading up to the record and of the record itself must be monotonically increasing, (2) the valid flag must be set, and (3) the checksum of the payload must match. The iterator ends when any one of these checks fails. These integrity checks guarantee that any partially written, uncommitted, or corrupted records are not applied during recovery.

4.4 Force Policy

Up until now, we have assumed that log writers would always do an explicit force on the log after writing a log record, i.e. that the thread doing the logging would not be able to continue until the log record was safely committed. But doing a force operation on the log after every write is expensive. The high cost of persisting data and the coarse granularity of writes for traditional block storage motivated researchers to propose techniques to reduce the number of synchronous, small writes required by relaxing *freshness* requirements (how up to date the state is after crash recovery). One popular approach is called group commit [11]. In this approach, log records are persisted in batches (or windows), such that the number of unpersisted records at any time is limited by a threshold,

³When configured without local PMEM, reserve returns a pointer to a buffer in DRAM.



Figure 4: Example showing the worst case scenario for the frequency-based force policy with $T=3$ and $F=2$, where all threads are blocked in trying to force persistence. All 6 records have been completed and may be lost on crash.

often called the group commit size. By controlling the group commit size, the desired level of freshness can be achieved.

With PMEM, the performance characteristics of persistent writes have changed drastically. PMEM write latency and granularity is much lower than other PCIe connected storage devices, so the cost of persistence is lower too. This seems to imply that there is no merit in limiting write frequency with PMEM. However, we find that when adding replication overhead to overall performance characteristics, the overhead of persistence remains relatively high. This is because the cost of replication over RDMA is often an order of magnitude higher than local writes to PMEM. To reduce the cost of replication, we first explore if group commit can be a viable approach. We find that although it improves overall latency, it is not scalable; at high concurrency, group commit has significant overhead. This is because a shared counter or mutex is required to maintain the current window. Concurrent access to this counter/mutex across threads results in significant cache thrashing and reduces the effectiveness of group commit. This effect was not observed in prior work either because they did not allow concurrent access to the log, or because the IO latency overshadowed synchronization overhead. With fast PMEM hardware, the software and synchronization overhead constitutes a significant portion of overall latency, so the effects of synchronization are much more pronounced.

To reduce the overhead of replication and persistence without adding synchronization overhead, we propose a novel frequency-based force policy. The key idea is as follows. Instead of maintaining a window of unpersisted records, we force records at intervals determined by a predefined ‘update frequency’. The crux of the idea lies in the fact that we can leverage the monotonicity of record LSNs to determine when to force data. For instance, given a frequency of F , data will be persisted every time a record with an LSN divisible by F is forced. Each thread that completes a record with such an LSN becomes the ‘force leader’ for the current batch of unpersisted records and is responsible for persisting them, which distributes this effort over all active threads. In this manner, we do not need a shared counter or mutex to maintain the window, but instead piggy back on existing synchronization to generate monotonic LSNs (in reserve) as consecutive integers.

In our design, threads can write and force records to the log

concurrently. Though the writes never overlap each other in the log space, they do race each other in time, and as a result, holes can be left in the log if a crash occurs after a thread writing further ahead in the log has completed its write while other threads haven’t yet filled in the records leading up to it. In order to provide total ordering in Arcadia, a force operation cannot return success until all records leading up to the forced record have been persisted and marked complete, thus a force later in the log must wait for any earlier force operations.

In the worst case, all threads can get blocked doing a force. This can happen if, for instance, the first thread doing a force is context switched, or if a record with an earlier LSN that is not a multiple of F has not yet completed. When threads force with the configured frequency, they are each separated by F records. This gives us the theoretical upper bound of the number of completed records that can be lost as $F \times T$, where T is the maximum number of threads or concurrent forces allowed. Figure 4 shows an example of the worst case scenario with $T=3$ and $F=2$. All three threads are blocked in this scenario. It is easy to see that the maximum number of records that may be lost on crash⁴ can be correctly calculated using the formula described here.

Use Cases. There are two use cases for this policy. The first is where explicit guarantees of persistence are not required for ensuring consistency (e.g., when recording logical operations in the log for a volatile key-value store). In this case, the frequency-based force policy can be used for all log updates to reduce the overhead of persisting records. A few of the recent updates may be lost on crash (the policy provides a bounded guarantee for how many updates can be lost), but consistency is not compromised. The second use case is where explicit guarantees of persistence are required for ensuring consistency (e.g., when recording state updates in the log for a database). Even in this case, the frequency-based force policy can be used for all log updates to reduce the overhead of persistence (by calling force with $\text{freq}=F$ for every record). When users require an explicit guarantee of record persistence (e.g., during the commit of a transaction), they may either issue a synchronous force (a force with $\text{freq}=1$) or check if the difference between the LSN of a newly allocated record and the record they want forced is greater than the theoretical upper bound of the size of the vulnerability window, $F \times T$.

5 Experimental Analysis

In this section, we present the evaluation of Arcadia. We conduct a comprehensive analysis of the impact of the novel aspects of our design. We also compare performance and resilience of Arcadia with state-of-the-art PMEM-optimized logs, FLEX [43], PMDK’s libpmemlog [17] (referred to as simply PMDK in the evaluation), and Query Fresh [42]. We were unable to compare with Tailwind [36] because it requires special hardware support and its source code is unavailable.

⁴We call this the vulnerability window.

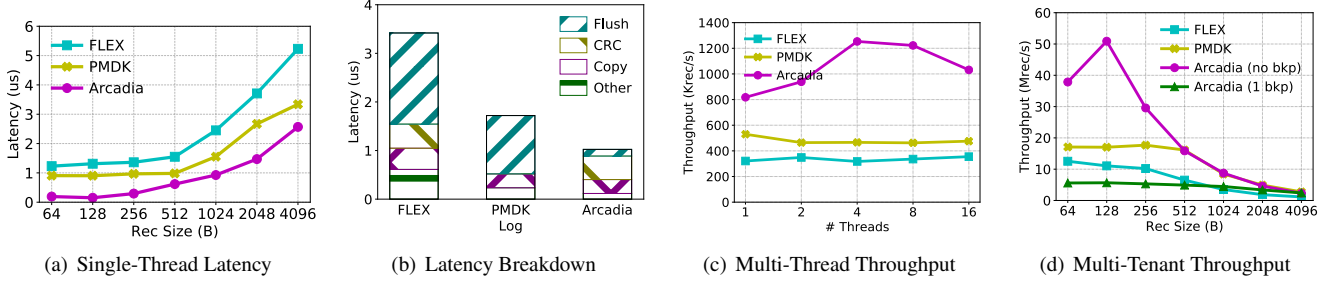


Figure 5: Microbenchmark Comparison with FLEX and PMDK

5.1 Experimental Testbed

Our experimental testbed consists of two Linux (4.18) nodes, each equipped with two Cascade Lake CPUs (5218@2.30GHz), 192GB DRAM (2 x 6 x 16GB DDR4 DIMMs), and 1.5TB PMEM (2 x 6 x 128GB DCPMMs) configured in App Direct mode. Each CPU has 16 physical cores (with hyperthreading enabled) and 22MB of L3 cache (LLC). All available PMEM is formatted as two ext4 DAX filesystems, each utilizing the memory of a single CPU socket as a single interleaved namespace. The two nodes are connected using 100Gbps EDR InfiniBand. One node is used as the primary while the other is used as a backup.

All code was compiled using GCC 8.3.1. PMDK [17] 1.8 was used across all evaluations to keep comparisons fair. Hardware counters were obtained using a combination of Intel VTune Profiler [16], Intel PCM [18], and Linux perf utility. DCPMM hardware counters were collected using the ipmwatch utility, a part of the VTune Profiler.

5.2 Microbenchmark Evaluation

We first present a microbenchmark-level comparison of Arcadia with FLEX and PMDK. Since FLEX and PMDK do not support replication, we compare them with Arcadia deployed in local mode.

Latency: We evaluate the latency of log writes with a single thread while varying the record size. Figure 5(a) shows this evaluation. We find that Arcadia has the lowest latency out of all three designs, up to 6x faster than PMDK and 8x faster than FLEX. One of the main reasons for this trend is that Arcadia does not maintain the pointer to the tail (most recently added record) of the log in PMEM (superline) and thus does not need to update it on every record write. FLEX performs the worst because it has high software overhead and it appends the record header and payload in separate operations. To confirm the reasons for these observations, we analyzed the breakdown of log writes for a 1KB record in Figure 5(b). We can clearly observe that flushing data to PMEM takes much more time in FLEX and PMDK. This is because they both update the log tail for each write. The flush time is especially high because it also includes the additional store fence required to wait for data being copied into PMEM before the tail update. By

avoiding a tail update, Arcadia shows much lower latency despite computing checksums.

Throughput: Figure 5(c) presents throughput curves in a multi-threaded setting. We measure the overall log throughput while increasing the number of threads concurrently adding records to the log. Both FLEX and PMDK have flat curves because they fully isolate writers using coarse-grained locks and do not provide any concurrency. Arcadia only isolates steps that require serialization (reserve and force), and can provide some level of concurrency. Its throughput is maximum at 4 threads, but reduces slightly at higher concurrency. There are two reasons for this – (1) synchronization overhead in reserve and force, and (2) poor bandwidth for highly concurrent accesses to PMEM. We also measure the aggregate log throughput in a multi-tenant setting. This is a common usage scenario where multiple tenants are sharing the same resources. Figure 5(d) shows the throughput for different record sizes with 16 concurrent single-threaded tenants, each writing to a separate log. Without replication, Arcadia performs the best in all cases. As the record size increases, throughput is bound by the PMEM bandwidth, and all log implementations converge to that limit. An observation we make is that Arcadia’s throughput is lower for 64B records than 128B records. This is because of the write amplification effects in DCPMM for small-sized writes. With replication, Arcadia has lower throughput than FLEX and PMDK for small record sizes. In these cases, replication overhead bounds overall throughput. However, for large record sizes (> 2KB), throughput is indistinguishable from the local mode.

5.3 Replication Overhead Analysis

To understand the performance characteristics of replication with PMEM and RDMA, we conduct an experiment to measure the replication overhead with Arcadia deployed in local+remote mode. To highlight performance impact, we measured three methods of write flush ordering: (1) parallel, in which the local cache flush and RDMA replication are done in parallel, (2) serial with the local flush performed first followed by the remote flush (LF+Rep.), and (3) serial with the remote flush done first followed by the local flush (Rep.+LF). Figure 6(a) analyzes the log write latency across record sizes for

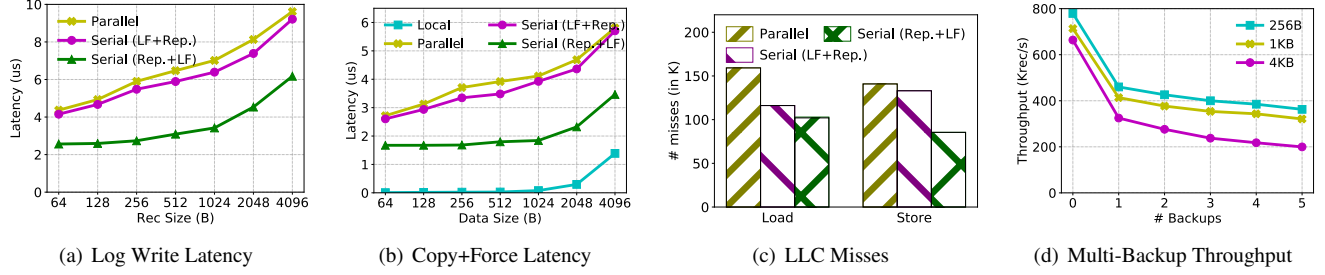


Figure 6: Replication Overhead Analysis

the three methods. Surprisingly, we find that the flush order has significant impact on overall latency. Contrary to expectation, parallel has the worst latency, while serial with remote flush first has the best. To understand why, we analyzed the overhead of just the copy and force operations of log writes (shown in Figure 6(b)) as well as the LLC miss counts of the three methods (shown in Figure 6(c)). The copy+force overhead curves mirror the log write latency curves, confirming that the local flush and fence operations are responsible for higher RDMA latency. Local flush is an order of magnitude faster than remote flush; therefore, when using the parallel method, the local cache flush invalidates data from the LLC.⁵ So, RDMA writes need to read data back from PMEM instead of reading it directly from the LLC. This additional read is manifested as additional LLC misses and is confirmed by Figure 6(c). The serial method with local flush first suffers from the same problem, and shows only marginally better performance compared to the parallel method. We suspect that this is because concurrent reads and writes to PMEM in the parallel method conflict and slightly reduce performance. The serial method with remote flushed first (Rep.+LF) has the best performance because RDMA writes can go data directly from the LLC — the LLC misses graph confirms this conclusion.

Next, we analyze the impact of number of log backups on throughput. Our testbed has only two nodes, so we conduct this experiment on another cluster with six Skylake nodes and EDR InfiniBand⁶. Figure 6(d) shows the impact of number of backups on throughput for different record sizes. A common observation across record sizes is that adding backups results in a significant drop in throughput. This is expected because the overhead of replication is significant. However, adding additional backups after the first one has little additional impact because the asynchronous nature of RDMA enables data to be sent to multiple backups in parallel. This is a significant result because it shows that whenever replication is enabled, throughput is not significantly impacted by the number of backups. Consequently, it's not necessary to compromise fault-tolerance for performance when replication is

⁵At the time of writing this paper, cache write-back instructions had not been implemented in Intel processors.

⁶This cluster does not have real PMEM, so we emulate PMEM with DRAM. This is also why we do not use this cluster for all other tests.

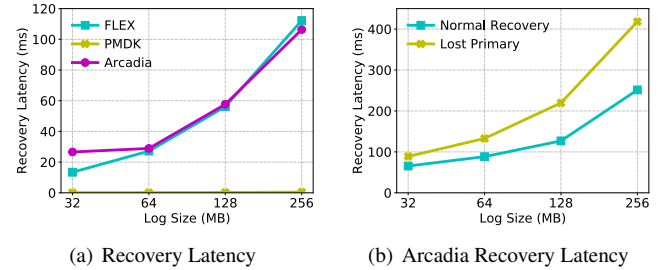


Figure 7: Recovery Evaluation

enabled. These results also show that in cases where PMEM is not available on all nodes, even nodes without PMEM can use Arcadia efficiently by having only remote log replicas.

5.4 Recovery Evaluation

To evaluate the recovery performance of Arcadia, we measured the total time taken to recover log state and iterate over all valid records on recovery, calling a null recovery function for each record. Figure 7(a) compares Arcadia's recovery latency in local mode with FLEX and PMDK. Both Arcadia and FLEX rely on checksums to verify record integrity, so they have similar performance. This is because the time taken to verify checksums dominates recovery latency, which is also why latency increases linearly with log size. PMDK does not use checksums, so its recovery procedure only consists of calling null functions for each record. This is why it is able to recover so fast. However, by not relying on any integrity checks, it is unable to handle undetected media errors and may redo or undo updates using corrupted data. Figure 7(b) compares Arcadia's recovery latency with replication enabled in normal recovery and when the primary log copy fails or is lost. These represent the best and worst case situations for recovery, respectively. We observe that when the primary copy is lost, the latency is higher. In this scenario, Arcadia needs to recover the primary copy using data in the backup which increases recovery time. However, this process uses the fast one-sided RDMA reads for copying data which does not add significant overhead. Even for a 256MB log, recovery takes less than 500ms in the worst case; so, we conclude that recovery is fast enough for most practical scenarios.

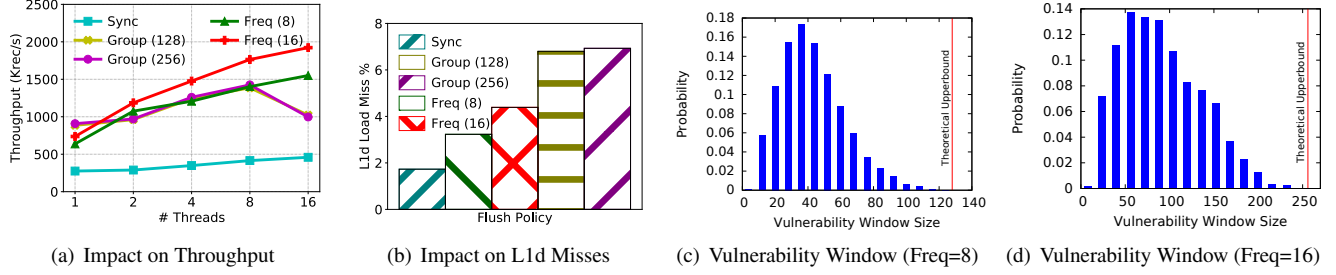


Figure 8: Force Policy Analysis

5.5 Force Policy Analysis

We compare group commit and frequency-based force policies in this experiment. We compare log throughput for the two policies in Figure 8(a). The group size or frequency is listed in parenthesis. We evaluate with two different values for each policy; these values have been chosen such that their theoretical vulnerability window sizes are comparable. So, group (128) is comparable to freq (8), while group (256) is comparable to freq (16). We also evaluate the sync policy in which each log write is synchronously forced. Results demonstrate the group commit has significant overhead at high concurrency and its throughput drops by 30% at 16 threads. Concurrent access to the window size variable is behind this drop. On the other hand, the frequency-based policy is much more scalable because it avoids this synchronization overhead. To confirm that synchronization overhead is the reason for poor performance of group commit, we analyze the L1d miss rate for all policies (see Figure 8(b)). We can clearly observe that group commit has much higher miss rates because of constant cache thrashing as a result of concurrent accesses to the shared counter.

We also measure the distribution of the vulnerability window size (from latest completed record to the most recently forced record) for the frequency-based policy. Unlike group commit, the vulnerability window size is not fixed because if a thread gets blocked doing a force, other threads can keep adding new records and increase the window size. Figures 8(c) and 8(d) show this distribution for frequencies of 8 and 16, respectively. Interestingly, we find that the probability distribution is skewed towards smaller sizes, far below the theoretical upper bound. This shows that on average, threads are unlikely to block on a force. Overall, we conclude that the frequency-based policy is not only more scalable, but in practice it provides better resilience than its theoretical limit.

5.6 Arcadia Applications

RocksDB. To demonstrate the practical benefits of Arcadia, we integrate it with RocksDB (adding/changing only ~200 LoC), a popular key-value database used at Facebook, by swapping out its write-ahead-log with Arcadia and use the fine-grained interface for adding log records. We evaluate

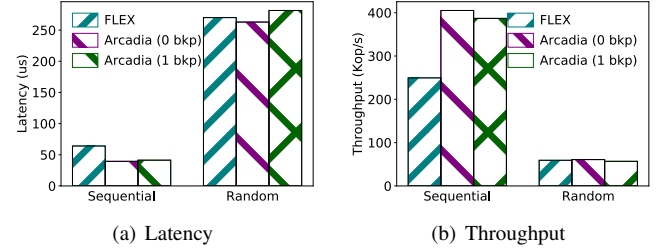


Figure 9: Comparison with FLEX using RocksDB

this version in both local and local+remote modes and compare it with the FLEX integration of RocksDB [32]. Figure 9 compares the latency and throughput for both sequential and random key-value puts at full-subscription (16 threads). By reducing the log append latency and using the fine-grained API to increase concurrency, Arcadia improves latency by up to 38% and throughput by up to 62% in local mode (0 bkp). In local+remote mode (1 bkp), Arcadia is faster than FLEX (which is local-only) for sequential puts and on-par for random puts, despite enabling replication. Replication overhead is much less than the overhead of the entire put operation. Hence, there is little difference in performance with and without replication. Finally, random puts are much slower than sequential puts, so latency is dominated by operations other than logging, and log performance differences are negligible.

Masstree. We integrate Arcadia with Masstree [28] (adding/changing only ~100 LoC), another popular key-value database, to enable comparison with Query Fresh. This experiment shows the practical benefits of the frequency-based force policy. Figure 10(a) compares the throughput of read-modify-writes and Figure 10(b) compares the theoretical vulnerability window size for Query Fresh and Arcadia (with both group commit and frequency-based force policies). From the results, we observe that Arcadia’s group commit policy has high synchronization overhead at 8 and 16 threads which reduces performance bringing it close to Query Fresh’s throughput. Query Fresh also uses group commit but it only enables limited log concurrency. Therefore, it delivers lower throughput than Arcadia but is also less impacted by the synchronization overheads of group commit. The frequency-based policy is

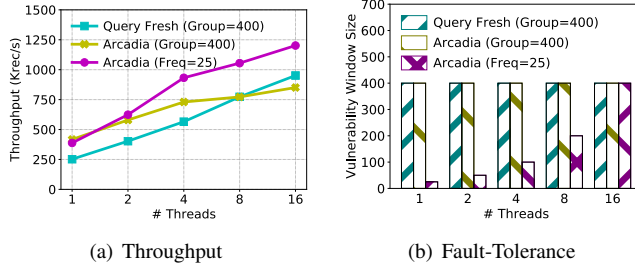


Figure 10: Comparison with Query Fresh using Masstree

able to deliver the best performance (up to 65% faster than Query Fresh) and fault-tolerance by avoiding unnecessary synchronization and forcing records more frequently.

5.7 Key Insights

Our measurements demonstrate the performance benefits of our approach compared to PMDK, FLEX, and Query Fresh. The measurements confirm the benefits of specific innovations and design choices in Arcadia, such as avoiding the superline tail pointer, and using concurrency to limit the impact of checksums and replication. They also highlight subtle factors such as the synchronization implied by group commit and the interaction between local flush and remote replication, giving confidence in the overall conclusions.

Arcadia’s log interface crucially decouples steps requiring serialization from those allowing concurrency to maximize performance of parallel applications. It also illustrates PMEM’s byte-addressability to avoid unnecessary data copying. The proposed frequency-based force policy gives flexibility in the freshness and performance trade-off, while allowing more concurrency than the traditional group commit method.

6 Related Work

Logging on PMEM. A number of prior works [4, 13, 17, 22, 36, 42, 43], have leveraged PMEM performance for logging as a means to improve overall system performance. Some of these works (Query Fresh [42], WBL [4], NV-Logging [13]) propose PMEM-based logs that are tightly integrated with a particular system or logging algorithm. Some of them (PMDK [17], FLEX [43], NVWAL [22]) propose unreplicated logs that cannot provide high availability. Others (e.g., Tailwind [36]) rely on special hardware support. Moreover, none of these are able to satisfy the robustness requirements of large-scale production systems because they do not provide resilience to all likely failure scenarios.

Logging on Flash/Disk. Corfu [6], Raft-based log [1], Apache Kafka [41], DistributedLog [9], GNR [7], and JBD2 [40] are relevant log implementations on traditional flash and disk storage that are deployed in production systems. These designs are considered to be reliable and robust. However, they were designed considering the performance charac-

teristics, access methods, and reliability of flash/disk which are considerably different than those of PMEM. Naïvely applying them to PMEM is unlikely to yield good results.

Combining RDMA and PMEM. Octopus [26], Forca [12], File-MR [45], Orion [44], pDPM [38], AsymNVM [27], NVFS [20], Hotpot [33], Mojim [48], and RDMP-KV [25] have worked on combining PMEM with RDMA for fast access to remote persistent storage. Each takes a different approach to remote persistence and atomicity that are not broadly applicable in other situations because they are either tied to a particular data format (key-value or file) or the complications of RDMA + PMEM persistence are not fully solved. In contrast, in this paper, we present abstract primitives for ensuring remote persistence, atomicity, and integrity, that can be applied generically on commodity hardware.

7 Conclusion

We’ve described Arcadia, a generic log with easy to use interface, stored on PMEM, and replicated using RDMA. Implementing fast, fault-tolerant systems is a challenging problem, and a good log implementation is a crucial abstraction for success. Arcadia efficiently encapsulates the operations of PMEM and RDMA to manage atomicity, persistence, and replication. It takes a creative approach in its API to eliminate unnecessary synchronization to provide the concurrency necessary to get the full performance from PMEM. We hope that wide adoption of this technology will significantly advance the frontier of fast and robust storage systems.

References

- [1] Jung-Sang Ahn, Woon-Hak Kang, Kun Ren, Guogen Zhang, and Sami Ben-Romdhane. Designing an Efficient Replicated Log Store with Consensus Protocol. In *HotCloud*, 2019.
- [2] Hiroyuki Akinaga and Hisashi Shima. Resistive Random Access Memory (ReRAM) Based on Metal Oxides. *Proceedings of the IEEE*, 98:2237–2251, 2010.
- [3] Apache ZooKeeper. <http://www.zookeeper.apache.org> (accessed Sep. 2021).
- [4] Joy Arulraj, Matthew Perron, and Andrew Pavlo. Write-Behind Logging. *Proceedings of the VLDB Endowment*, 10(4):337–348, 2016.
- [5] InfiniBand Trade Association. Supplement to Infiniband Architecture Specification Volume 1, Release 1.2. 1: Annex A17: RDMA over Converged Ethernet (RoCEv2), 2014.
- [6] Mahesh Balakrishnan, Dahlia Malkhi, Vijayan Prabhakaran, Ted Wobblers, Michael Wei, and John D Davis.

- CORFU: A Shared Log Design for Flash Clusters. In *NSDI*, pages 1–14, 2012.
- [7] Veera Deenadhayalan. GPFS Native RAID for 100,000-Disk Petascale Systems. In *LISA*, 2011.
- [8] Chet Douglas. RDMA with PMEM. http://www.snia.org/sites/default/files/SDC15_presentations/persistent_mem/ChetDouglas_RDMA_with_PM.pdf, 2015.
- [9] Sijie Guo, Robin Dhamankar, and Leigh Stewart. DistributedLog: A High Performance Replicated Log Service. In *ICDE*, pages 1183–1194, 2017.
- [10] Frank T Hady, Annie Foong, Bryan Veal, and Dan Williams. Platform Storage Performance with 3D XPoint Technology. *Proceedings of the IEEE*, 105(9):1822–1833, 2017.
- [11] Pat Helland, Harald Sammer, Jim Lyon, Richard Carr, Phil Garrett, and Andreas Reuter. Group Commit Timers and High Volume Transaction Systems. In *International Workshop on High Performance Transaction Systems*, pages 301–329, 1987.
- [12] Haixin Huang, Kaixin Huang, Litong You, and Linpeng Huang. Forca: Fast and Atomic Remote Direct Access to Persistent Memory. In *ICCD*, pages 246–249, 2018.
- [13] Jian Huang, Karsten Schwan, and Moinuddin K Qureshi. NVRAM-aware Logging in Transaction Systems. *Proceedings of the VLDB Endowment*, 8(4):389–400, 2014.
- [14] InfiniBand Trade Association. <http://www.infinibandta.org> (accessed Sep. 2021), 1999.
- [15] Intel. PMEM Write Atomicity. <https://software.intel.com/content/www/us/en/develop/articles/persistent-memory-replication-over-traditional-rdma-part-1-understanding-remote-persistent.html#inpage-nav-4-2> (accessed Sep. 2021).
- [16] Intel. Intel VTune Profiler. <https://software.intel.com/content/www/us/en/develop/tools/vtune-profiler.html> (accessed Sep. 2021), 2014.
- [17] Intel. PMDK. <https://github.com/pmem/pmdk> (accessed Sep. 2021), 2014.
- [18] Intel. Processor Counter Monitor. <https://github.com/opcm/pcm> (accessed Sep. 2021), 2017.
- [19] Intel. 300 nanoseconds. <https://pmem.io/2019/12/19/performance.html> (accessed Sep. 2021), 2019.
- [20] Nusrat Sharmin Islam, Md Wasi-ur Rahman, Xiaoyi Lu, and Dhabaleswar K Panda. High Performance Design for HDFS with Byte-Addressability of NVM and RDMA. In *ICS*, page 8, 2016.
- [21] Hyojun Kim, Sangeetha Seshadri, Clement L Dickey, and Lawrence Chiu. Evaluating Phase Change Memory for Enterprise Storage Systems: A Study of Caching and Tiering Approaches. In *FAST*, pages 33–45, 2014.
- [22] Wook-Hee Kim, Jinwoong Kim, Woongki Baek, Beom-seok Nam, and Youjip Won. NVWAL: Exploiting NVRAM in Write-Ahead Logging. In *ASPLOS*, pages 385–398, 2016.
- [23] Benjamin C Lee, Engin Ipek, Onur Mutlu, and Doug Burger. Architecting Phase Change Memory as a Scalable DRAM Alternative. In *ISCA*, pages 2–13, 2009.
- [24] Benjamin C Lee, Ping Zhou, Jun Yang, Youtao Zhang, Bo Zhao, Engin Ipek, Onur Mutlu, and Doug Burger. Phase-Change Technology and the Future of Main Memory. *IEEE Micro*, 30(1):143, 2010.
- [25] Tianxi Li, Dipti Shankar, Shashank Gugnani, and Xiaoyi Lu. RDMP-KV: Designing Remote Direct Memory Persistence based Key-Value Stores with PMEM. In *SC*, pages 722–735, 2020.
- [26] Youyou Lu, Jiwu Shu, Youmin Chen, and Tao Li. Octopus: An RDMA-enabled Distributed Persistent Memory File System. In *USENIX ATC*, pages 773–785, 2017.
- [27] Teng Ma, Mingxing Zhang, Kang Chen, Zhuo Song, Yongwei Wu, and Xuehai Qian. AsymNVM: An Efficient Framework for Implementing Persistent Data Structures on Asymmetric NVM Architecture. In *ASPLOS*, pages 757–773, 2020.
- [28] Yandong Mao, Eddie Kohler, and Robert Tappan Morris. Cache Craftiness for Fast Multicore Key-Value Storage. In *EuroSys*, pages 183–196, 2012.
- [29] Sparsh Mittal and Jeffrey S Vetter. A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems. *IEEE Transactions on Parallel and Distributed Systems*, 27(5):1537–1550, 2016.
- [30] C. Mohan, Donald J. Haderle, Bruce G. Lindsay, Hamid Pirahesh, and Peter M. Schwarz. ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging. *ACM Transactions on Database Systems (TODS)*, 17:94–162, 1992.
- [31] Sanketh Nalli, Swapnil Haria, Mark D Hill, Michael M Swift, Haris Volos, and Kimberly Keeton. An Analysis

- of Persistent Memory use with WHISPER. In *ASPLOS*, pages 135–148, 2017.
- [32] Peifeng Si. Persistent Memory Storage Engine for RocksDB. <https://github.com/pmem/pmem-rocksdb> (accessed Sep. 2021), 2019.
- [33] Yizhou Shan, Shin-Yeh Tsai, and Yiyang Zhang. Distributed Shared Persistent Memory. In *SoCC*, pages 323–337, 2017.
- [34] Steve Swanson. Engineering Reliable Persistence. <https://www.sigarch.org/engineering-reliable-persistence/> (accessed Sep. 2021), 2017.
- [35] Dmitri B Strukov, Gregory S Snider, Duncan R Stewart, and R Stanley Williams. The Missing Memristor Found. *Nature*, 453(7191):80, 2008.
- [36] Yacine Taleb, Ryan Stutsman, Gabriel Antoniu, and Toni Cortes. Tailwind: Fast and Atomic RDMA-based Replication. In *USENIX ATC*, pages 851–863, 2018.
- [37] Tom Talpey. Remote Access to Ultra-Low-Latency Storage - SNIA. https://www.snia.org/sites/default/files/SDC15_presentations/persistent_mem/Talpey-Remote_Access_Storage.pdf, 2015.
- [38] Shin-Yeh Tsai, Yizhou Shan, and Yiyang Zhang. Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores. In *USENIX ATC*, pages 33–48, 2020.
- [39] AA Tulapurkar, Y Suzuki, A Fukushima, H Kubota, H Maehara, K Tsunekawa, DD Djayaprawira, N Watanabe, and S Yuasa. Spin-Torque Diode Effect in Magnetic Tunnel Junctions. *Nature*, 438(7066):339, 2005.
- [40] Stephen C Tweedie. Journaling the Linux ext2fs Filesystem. In *The Fourth Annual Linux Expo*, 1998.
- [41] Guozhang Wang, Joel Koshy, Sriram Subramanian, Kartik Paramasivam, Mammad Zadeh, Neha Narkhede, Jun Rao, Jay Kreps, and Joe Stein. Building a Replicated Logging System with Apache Kafka. *Proceedings of the VLDB Endowment*, 8(12):1654–1655, 2015.
- [42] Tianzheng Wang, Ryan Johnson, and Ippokratis Pandis. Query Fresh: Log Shipping on Steroids. *Proceedings of the VLDB Endowment*, 11(4):406–419, 2017.
- [43] Jian Xu, Juno Kim, Amirsaman Memaripour, and Steven Swanson. Finding and Fixing Performance Pathologies in Persistent Memory Software Stacks. In *ASPLOS*, pages 427–439, 2019.
- [44] Jian Yang, Joseph Izraelevitz, and Steven Swanson. Orion: A Distributed File System for Non-Volatile Main Memory and RDMA-Capable Networks. In *FAST*, pages 221–234, 2019.
- [45] Jian Yang, Joseph Izraelevitz, and Steven Swanson. FileMR: Rethinking RDMA Networking for Scalable Persistent Memory. In *NSDI*, pages 111–125, 2020.
- [46] Jian Yang, Juno Kim, Morteza Hoseinzadeh, Joseph Izraelevitz, and Steve Swanson. An Empirical Guide to the Behavior and Use of Scalable Persistent Memory. In *FAST*, pages 169–182, 2020.
- [47] Lu Zhang and Steven Swanson. Pangolin: A Fault-Tolerant Persistent Memory Programming Library. In *USENIX ATC*, pages 897–912, 2019.
- [48] Yiyang Zhang, Jian Yang, Amirsaman Memaripour, and Steven Swanson. Mojim: A Reliable and Highly-Available Non-Volatile Memory System. In *ASPLOS*, pages 3–18, 2015.