

# A High-Performance Persistent Memory Key-Value Store with Near-Memory Compute

Daniel G. Waddington  
*daniel.waddington@ibm.com*  
IBM Research Almaden

Clem Dickey  
*dickeycl@us.ibm.com*  
IBM Research Almaden

Luna Xu  
*xuluna@ibm.com*  
IBM Research Almaden

Moshik Hershcovitch  
*moshikh@il.ibm.com*  
IBM Research Almaden

Sangeetha Seshadri  
*seshadrs@us.ibm.com*  
IBM Research Almaden

## Abstract

MCAS (Memory Centric Active Storage) is a persistent memory tier for high-performance durable data storage. It is designed from the ground-up to provide a key-value capability with low-latency guarantees and data durability through memory persistence and replication.

To reduce data movement and make further gains in performance, we provide support for user-defined “push-down” operations (known as Active Data Objects) that can execute directly and safely on the value-memory associated with one or more keys. The ADO mechanism allows complex pointer-based dynamic data structures (e.g., trees) to be stored and operated on in persistent memory.

To this end, we examine a real-world use case for MCAS-ADO in the handling of enterprise storage system metadata for Continuous Data Protection (CDP). This requires continuously updated complex metadata that must be kept consistent and durable.

In this paper, we i.) present the MCAS-ADO system architecture, ii.) show how the CDP use case is implemented, and finally iii.) give an evaluation of system performance in the context of this use case.

## 1 Introduction

While the notion of persistent memory has been around for many decades, its widespread availability had not arisen until recently. By 2021, Intel’s Optane 3D-Xpoint Persistent Memory (PM) DIMM market is projected to reach \$1B in revenue [3]. Its value proposition, compared to DRAM, is centered around lower-cost per GB, larger capacity, lower-energy consumption through refresh elimination, and data retention (persistence) in power-failure and reset events.

PM is *byte-addressable* in that it is directly accessed via load-store instructions of the CPU. The current generation of Intel’s Optane PM operates at a read-write latency of around 300ns [7, 16]. Even though this is slower than DRAM access latencies ( $\sim 100ns$ ) it is at least 30x faster than state-of-the-art storage (e.g., NVMe SSD). Capacity of PM is also about 8x that of DRAM<sup>1</sup>.

As with storage, in order to enable data sharing and availability PM must also be provisioned as a network-attached resource. Today, a growing number of key-value stores use memory for performance and replication for reliability [12, 18, 19, 21, 22, 24, 27, 30]. More recently, key-value stores that support PM are emerging in academia [10, 17, 32]. However, this existing

<sup>1</sup>At least for 3D-XPoint, which is based on lattice-arranged Phase Change Memory (PCM).

work is limited in its ability to take full advantage of near-data compute.

The key contributions of the paper are as follows:

- i. Extensions of the “plain-old” network-attached MCKVS key-value store [33] that provide flexible Active Data Object (ADO) capabilities allowing arbitrary compute to be safely *pushed down* into the storage system.
- ii. Demonstration of the solution in a real-world use case based on a fast and durable indexing data structure for Continuous Data Protection (CDP) in enterprise storage systems.
- iii. Evaluation of general throughput and latency performance/scaling.
- iv. Comparison of performance for the ADO paradigm versus a plain-old key-value paradigm indicating a 43% performance improvement in our given use case.

The rest of the paper is organized as follows: Section 2 provides a background on MCAS, which is derived from MCKVS [33], and then describes the architecture of Active Data Objects (ADO) extension. Section 3 details the application of ADO extensions to various functions within an over-arching CDP index service. Section 4 presents the evaluation results. Section 5 and 6 summarize the related work and conclusions respectively.

## 2 MCAS Design

At a fundamental level, MCAS is a network-attached key-value store that is enhanced to allow flexible in-store operations directly on persistent memory. It is a *sharded* architecture enabling a lock-free design with the restriction of only allowing 1 :  $N$  shard-to-pool mapping (i.e. any pool can only be serviced by a single shard). Each shard is associated with a set of persistent memory regions and CPU resources. Memory

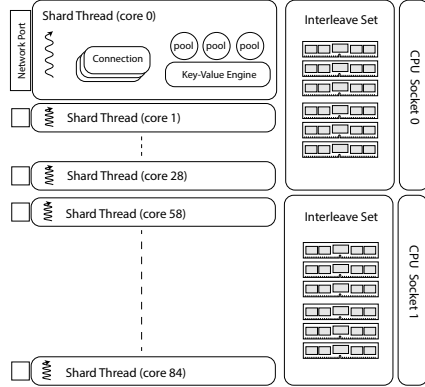


Figure 1: MCAS Sharding Architecture

regions can be either a device DAX (devdax) partition or file-system DAX (fsdax) file<sup>2</sup>. At the socket level, Optane-PMM is configured in App-Direct mode and data is striped across six DIMMs in order to achieve maximum socket performance.

Each shard also maintains a single network connection end-point (port). MCAS uses *libfabric* with the RDMA verbs or TCP/IP sockets provider. For the work in this paper, the RDMA provider is used since this offers high-performance kernel-bypass access to the network.

Data is arranged in *pools* that represent a set of persistent memory resources, which are allocated from a coarse-grained allocator managing the shard’s memory regions. Pool memory is used for the hash table index, keys and values, and can be dynamically expanded. Each pool represents a logical set of key-value pairs in a single key namespace.

MCAS supports variable key and value sizes. This means that a heap-allocator is necessary to support variable-length region allocation. To support high-rates of key-value pair insertion and deletion, MCAS maintains a heap allocator for key-value data in volatile (DRAM) memory.

<sup>2</sup>Use of file-system DAX requires On-Demand Paging hardware in the RDMA NIC.

The core of the persistent key-value index is a hopscotch hash table [15], which maps keys to values. The hash table for a given key-value pool is maintained in persistent memory.

Further detail on MCAS core (MCKVS) is given in [33].

## 2.1 Active Data Objects

A key feature of MCAS is its ability to support in-store operations via Active Data Object (ADO) plugins. This allows the advanced user to associate in-store compute with a given pool.

ADOs are “sandboxed” compute that allow for generalized in-place execution on persistent memory that is managed by MCAS. They execute outside of the shard process, which maintains the key-value index, so that their scope of data access can be limited. ADOs enable the layering of both domain-specific (e.g., matrix multiply) and generalized services (e.g., replication, versioning) over the basic key-value store.

ADOs are implemented as plugins. The plugins are *invoked* via the following client-side operation:

```
status_t invoke_ado(
    const IMCAS::pool_t pool,
    const std::string& key,
    const void* request,
    const size_t request_len,
    const ado_flags_t flags,
    std::vector<ADO_response>& out_response,
    const size_t value_size = 0);
```

On instantiation, the shard process maps the regions of persistent memory corresponding to the pool into the ADO process. Thus, the ADO process has complete visibility of the pool memory (the security boundary is the pool).

The invoke operations incorporate an *opaque request*. This passes from the client, through the shard process and then into the ADO process via a user-level IPC connection. Ultimately, on receiving a request, the ADO process invokes the plugin’s `do_work` operation:

```
status_t do_work(
    const uint64_t work_id,
    const char* key,
    const size_t key_len,
    IADO_plugin::value_space_t& values,
    const void* in_work_request,
    const size_t in_work_request_len,
    const bool new_root,
    response_buffer_vector_t& response_buffers);
```

The `key` and `key_len` parameters are provided as a reference point because actions are normally logically associated with a given key-value pair. Nevertheless, the ADO plugin has full visibility of the key-value space within the whole pool. In order to interact with the memory and data therein, the ADO plugin uses call-backs to the shard process to manage memory, see Table 1.

Functions	Description
create/open/erase	Key-value management
resize value	Resize existing value
allocate/free memory	Pool memory management
get ref vector	Get vector of key-value pairs
iterate	Iterate key-value pairs
find key	Scan for key through secondary index
get pool info	Retrieve memory utilization etc.
unlock	Explicitly unlock key-value pair

Table 1: ADO plugin callback API

ADO plugins can also be layered. Signals and data can be exchanged between layers via the response buffer vector and/or shared memory of the pool (e.g., a special key-value pair). Layered plugins are invoked in a round-robin manner.

A subtle design characteristic of the ADO plugin architecture, is that only the plugin itself can interpret the opaque message. Thus, the ADO plugins and client-side libraries (that construct the requests) are packaged as part of a readily extensible *open protocol* (see Figure 2). The advantage of this design is that any arbitrary functionality can be implemented in the ADO plugin, e.g., replication, encryption, data structure manipulation, tiering. The opaque request can be program (e.g., compiled or interpretable

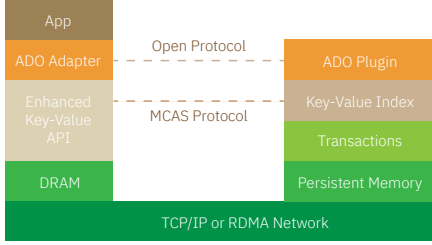


Figure 2: MCAS Protocol Stack

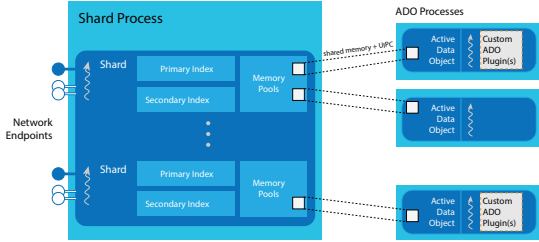


Figure 3: ADO Architecture

code) or data (e.g., encrypted message).

MCAS also supports an `invoke_put_ado` operation, which allows the client to “put” a value into the store prior to invoking the ADO plugin.

Any data that is operated on by the ADO plugin must be *crash-consistent* [25]. This means that if the system fails at any point of time, the data structure can at least be recovered to a *transactionally complete* state. The most common approach to realizing crash-consistency is to use an *undo-log* or *redo-log*.

Other more advanced techniques that rely on novel hardware are available that are less onerous on the software [8, 11, 26, 31]. The choice of approach to crash-consistency is not dictated by MCAS and is left up to the ADO plugin developer. In our use case implementation described later in Section 3, we use a modified C++ Standard Templates Library to provide transparent undo logging together with 64-bit atomic operations.

### 3 Applying the MCAS-ADO approach to a CDP Index

Ransomware has rapidly emerged as the most visible cybersecurity risk threatening business, government and private sectors. It is a class of malware that allows the attacker to encrypt data in order to then extort the victim into paying a ransom, normally through crypto-currency, for its decryption and restoration.

A key strategy in combating ransomware is to ensure that data is copied into a secure storage tier, thereby allowing data to be recovered in the event of attack. Continuous Data Protection (CDP) refers to transparently making a backup of data in real-time and thus maintaining an immutable history of the data which can be used for restoration while minimizing data loss [29].

The focus use case, for our evaluation of the MCAS-ADO approach, is achieving higher performance and scalability for a CDP index implementation. The broader CDP solution (outside the scope of this paper) is to combine a distributed block storage system with a high-performance, durable (replicated) index that maintains a mapping between *virtual blocks* and *managed blocks* over time (see Figure 4). Virtual blocks are those that the application “sees”, albeit through a filesystem in most cases. As with locally attached storage, the virtual space only represents the latest version.

Managed blocks collectively represent versions over time. When the application state needs to be rolled-back to a prior version during recovery, the virtual space for that point-in-time is reconstructed from the managed blocks.

In terms of normal behavior outside of an attack, the index must be updated for each write in the system. For example, 10,000 containerized applications issuing 200 IOPS each (typical for a DB application) requires 2M updates/second on the index.

The low-latency and guaranteed synchronous persistency provided by MCAS enables the index to be kept consistent with the storage system at fine granularity, while avoiding loss of

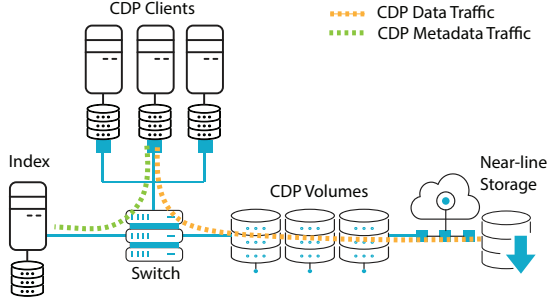


Figure 4: Basic CDP Architecture

any data in the event of power failure or reset event. Ultimately, this reduces complexity and recovery time by avoiding the need to reconstruct the complete or partial index in memory when the system starts up. Furthermore, the high update performance and smaller memory footprint, which is a key element of our design, makes scaling to thousands of applications possible.

**The CDP Index** As previously discussed, the basic premise is to map a set of virtual blocks to managed blocks over time.

Figure 5 illustrates a simple example where three writes create new mappings to the managed block space. The numbers in parenthesis designate the offset in the managed space for the given mapped virtual range. In the example, each new write is appended to the managed space. Mapped ranges can represent one or more blocks and they do not overlap.

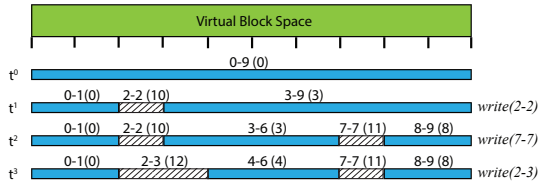


Figure 5: Example Block Mapping over Time

In order to scale to large containerized environments supporting thousands of persistent

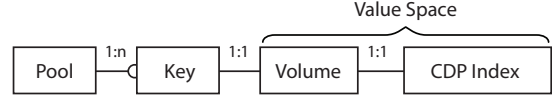


Figure 6: Data Schema

volumes, we implement the CDP use case using a 2-level index organization. At the top-level, each entry in a hash table represents a key-value pair. The key is a Volume Tag (a shortened version of a longer name) which uniquely identifies an application volume (see Figure 6). The value stores the handle to a persistent data structure representing the CDP Index for that volume. The CDP Index, described in Figure 7, records all mappings from virtual to managed blocks over time. Each time a block is modified a new entry is created in the CDP Index.

In order to support millions of updates per second, update operations must exhibit low latency and must be durable; lazy persistency or lazy replication lead to potential data loss and more recovery complexity. To support this performance requirement we leverage the flexibility of the ADO plugin mechanism, which allows arbitrary pointer-based data structures to be instantiated and modified.

From a logical point of view, the index maintains an up-to-date block map for every point-in-time. However, to reduce the overall memory footprint and compute demand, our design batches updates and performs materialization of the map only on completion of the batch. Each batch referred to as a `Time_quantum` contains updates that occur during a time range.

Each write adds a `Persistent_managed_range` 64-byte record<sup>3</sup> to the current `Time_quantum`.

Each record write is crash-consistent and made transactional through a combination of undo-logging and atomic operations.

<sup>3</sup>Being equal to the size of a cache line helps minimize read-write amplification, although some amplification still happens internally to the NVDIMM.

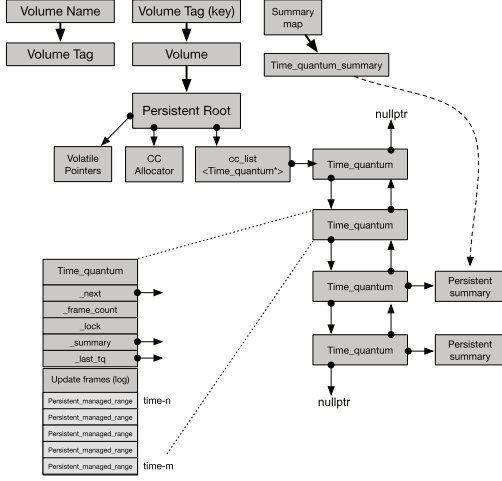


Figure 7: CDP ADO Data Structure

### 3.1 Lazy Summarization and Point-in-time Query

The memory size for each quantum is configurable. However, between 4MiB (64K records) and 16MiB (262K records) is typical. On reaching full capacity, each quantum is “handed over” to a secondary thread in the ADO plugin that performs *lazy summarization*. The summarization process builds the mapping state for the last point in time in the quantum (i.e. the time point of last record). This is achieved by taking the prior quantum’s summary (held in DRAM) and merging in all of the writes from the quantum in time order. The merge process is performed in DRAM and on completion transactionally copied and made durable in persistent memory.

Memory for the persistent data structure elements is allocated by a local heap in the ADO plugin. The heap is populated with coarse-grained regions of memory (chunks) to manage, that are allocated from the memory pool via a call-back to the shard process. The exact size of the chunks is configurable (64MiB is typical).

The fundamental purpose of the index is to provide a block mapping for a specific instance

in time,  $t$ . After an attack event, the CDP recovery process iteratively explores mappings for multiple points in time. Mapping requests may be for the full block-range or some narrower sub-range of the volume. The algorithm used to iteratively determine an effective recovery point is outside the scope of this paper.

Retrieval of a mapping for time  $t$  is as follows:

1. Starting from the most recent, scan each element in the quantum list until a `Time_quantum`  $Q_t$  is identified that has a timestamp less than or equal to  $t$ .
2. If the identified quantum does not have a summary (recall lazy summarization), continue searching backwards until a quantum with complete summary  $Q_s$  is identified ( $time(Q_s) \leq time(Q_t)$ ).
3. Starting with summary  $Q_s$  merge in all updates (`Persistent_manage_range` records), in time order, into a result  $R$ . Continue merging updates for quanta up to  $Q_t$ .
4. Merge all records  $r \in Q_t$  while  $time(r) \leq t$  into result  $R$ .

Because lazy summarization is able to keep up with the incoming updates,  $Q_t$  and  $Q_s$  are in most cases the same quantum. If the query is for a sub-range, then updates are filtered by that range (i.e. included if they overlap).

### 3.2 Aging Out Data

To guard against delayed attack detection, mapping data is retained for a period of time. As the data’s age extends beyond a pre-defined threshold (e.g. 7 days) the mapping data is copied to a slower tier and erased in MCAS. Our design uses a secondary ADO plugin thread to “trim” the `Time_quantum` list periodically. This is done by scanning the crash-consistent `Time_quantum` list (`cc_list`) and then removing the tail element should its timestamp indicate that it is sufficiently old. After removal from the list, the memory for the quantum and its associated



summary are released back to the ADO-local heap.

### 3.3 Data Replication

To protect against failures within a data center, our CDP use case also requires that multiple replicas of the mapping data exist.

The CDP plugin uses a client-side replication strategy. Achieving consistency requires that write-updates for a given volume are sent from a single client, which holds true in our broader design. The client does not send the next update until it has received an acknowledgement for the prior update from all nodes. Although details of failure handling, reconciliation and recovery are outside the scope of this paper, we include in the evaluation the write throughput using two- and three-way replication demonstrating the ability of MCAS to sustain high-performance even with synchronous replication.

## 4 Evaluation

In this section, we present our experimental results for the CDP use case. We examine performance of both non-replicated and replicated update (write) workload as well as sporadic point-in-time query latencies (read). In the current design, the client maintains a local cache with the latest block mappings. A majority of the reads are directly served out of the local cache. Hence a 100% query workload is not a useful measure for this use case. On the other hand, every write operation results in a write operation to the MCAS backend to update the index. As a result, write throughput (Section 4.1) is a key measure of performance.

In order to evaluate the benefit of near-data compute, we compare the CDP ADO performance with a “thick” client alternative that uses MCAS as a basic key-value store and performs CDP quantum summarization and query handling on the client node.

Table 2 details our experimental hardware and software configurations. Figure 8 shows

Component	Description
Processor	Intel Xeon Gold 5128 (Cascade Lake) 16-core 2.30GHz
Cache	L1 (32KiB), L2 (1MiB), L3 (22MiB)
DRAM	PC2400 DDR4 16GiB 12x DIMMs (192GiB)
NVDIMM	Optane-PMM 128GB 12x DIMMs (1.5TB)
RDMA NIC	Mellanox ConnectX-5 (100GbE)
OS	Linux Fedora 27 with Kernel 4.18.19 x86_64
Compiler	GCC 7.3.1
NIC S/W	Mellanox OFED 4.5-1.0.1

Table 2: Server system specification

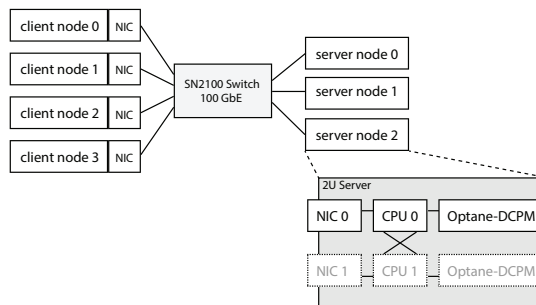


Figure 8: Experimental Network Topology

the network topology. Note that the scaling experiments use a single processor, with NUMA-local memory, and a single NIC.

Each pinned shard thread has a 1:1 association with a pinned ADO primary thread. A fraction of threads are kept floating to serve as ADO secondary threads. These are used to perform background operations such as summarization and age-out. A fully populated single-socket that supports 32 cores (using Hyper-Threading) will support up to 12 shards, 12 primary ADO threads and 8 secondary threads. A two-socket server can support twice this.

## 4.1 Index Write Performance

This experiment measures the maximum write throughput and scaling across an increasing number of shards. Clients generate random mapping updates for volumes spanning 1M blocks with a random span of 1-100 blocks. Update timestamping uses actual time (nanosecond epoch). Quantums are 64K records and each is aged out after a maximum of 10 newer quantums exist. Each client uses 1, 2 or 6 threads executing updates to different volumes. Note that 6 threads is sufficient to saturate performance on a single shard. The throughput mean is taken over a 10 minute period.

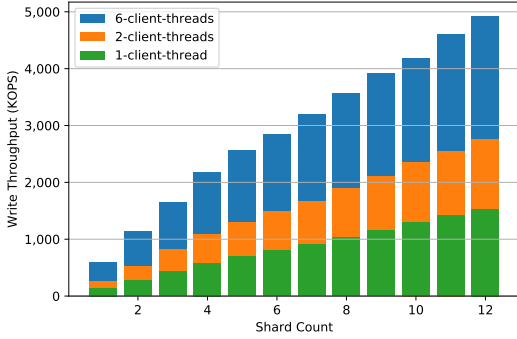


Figure 9: Throughput Scaling without Replication

For the update workload, Figure 9 shows throughput scaling of up to 4.92M updates/second, while a single client thread can achieve 140K updates/second with a mean round-trip time of  $7.14\mu\text{sec}$ .

**Client Observed Update Latency** Figure 10 shows latency distributions for a single shard operating with 1 and 6 client-side threads. A logarithmic scale is used for the Y-axis and outliers have not been removed. The data shows that 99.77% of the latency for the 1-threaded client is less than  $10\mu\text{sec}$ , while 99.39% of writes for the 6-threaded client are also less than  $10\mu\text{sec}$ . As would be expected, latency for

the 6-thread scenario is increased as threads compete in a queued fashion.

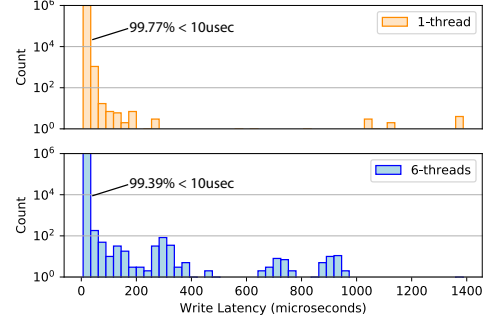


Figure 10: Round-Trip Latency Distribution

We believe that this tight long-tail latency cannot be achieved with other traditional storage technologies such as NVMe SSD.

**Replication Scaling** Using 6 threads per client, Figure 11 illustrates scaling for two- and three-way replication. For 12 shards, the system is able to scale to 3.92M and 3.07M updates/second for two- and three-way replication, respectively. This represents a degradation of 20% and 37%.

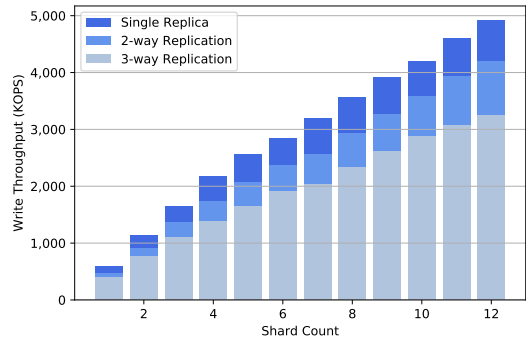


Figure 11: Throughput Scaling with Replication



## 4.2 Query Latency

We now explore the query latency characteristics of our solution. To re-cap, a query consists of materializing a virtual-to-managed block mapping at some point-in-time  $t$ . This is achieved by using the time  $t$ 's prior quantum summary as a starting point and then merging in updates up to time  $t$ . Thus, the time taken to materialize is dependent upon the size of the quantum and the position of  $t$  within the quantum.

To illustrate this, Figure 12 shows latencies for different quantum sizes: 4MiB (65K records), 8MiB (131K records) and 16MiB (262K records). Time points are randomly selected and the mapping for a contiguous region of 100K blocks is retrieved. The data shows that for each 4MiB increase in quantum size the worst case latency is increased by approximately 50ms.

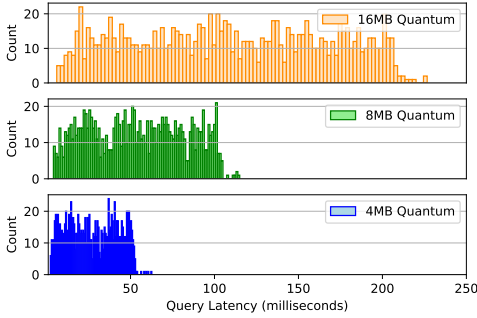


Figure 12: Query Latency Distributions for Different Quantum Sizes

**Query Under Load** We also examine how queries impact on-going concurrent write streams. In our use case, we expect that in most cases not all clients will need to recover together; some clients will continue to function as normal. To this end, we examine how throughput rate for a single client (which because of the synchronous behavior correlates directly with latency) is affected by another client making concurrent query requests on the same shard.

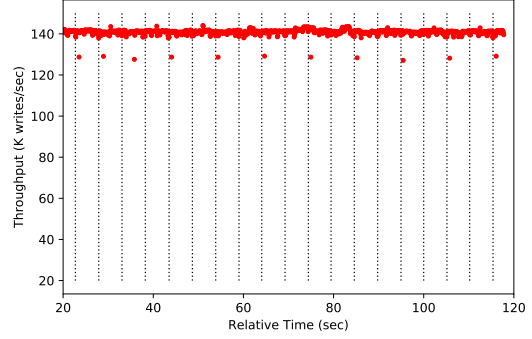


Figure 13: Query Latency Under Load

Given writing client  $C_w$  and query client  $C_q$ , Figure 13 shows how the throughput of  $C_w$  is impacted by periodic queries made by  $C_q$ . The vertical dotted lines mark the point-in-time at which query results are received by  $C_q$ . The red dots represent throughput samples at 12.5K updates intervals by  $C_w$ . The maximum throughput observed is 144K updates/second and the minimum observed is 127K updates/second. Worst-case degradation is thus 11.8%.

## 4.3 Comparison to Non-ADO Alternative Architecture

In order to better validate and quantify the performance improvement resulting from the MCAS ADO architecture, we compare the ADO-based solution with a “plain” key-value solution that performs the same CDP index operations but on the client side. To achieve a fair comparison, we use the basic MCAS key-value store without the ADO feature [33].

Figure 14 shows the Plain-KV deployment. The CDP data structures are broken down into individual key-value pairs. The primary key is the `QuantumId`, which is derived from the Volume Name and a unique numerical identifier.

Local copies of all data are maintained in the client’s local volatile memory (DRAM) and hence in practice would be limited to the available client memory. Copies of the update frames

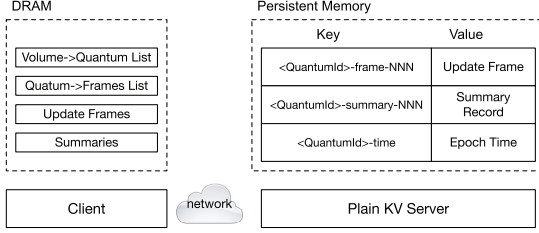


Figure 14: Alternative Deployment with Plain Key-Value Store

and any summaries are copied in the remote store’s persistent memory for durability. However, queries on the data are made from local copies alone (except when performing client-side rebuild). Data age-out is realized as explicit key-value pair deletions.

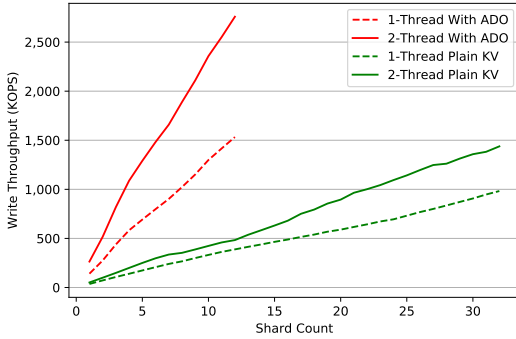


Figure 15: Comparison of Throughput for ADO vs. Plain-KV Deployments

**Throughput** Figure 15 shows a comparison of throughput under a 100% write/update workload.

The Plain-KV deployment does not need ADO threads and therefore is able to scale shards further to a maximum of 32 (at one thread per shard). The figure shows 1 and 2 client threads; we do not have sufficient test client machines available to provide data for 6 threads/client for Plain-KV. The data shows that the ADO-based deployment, with the same

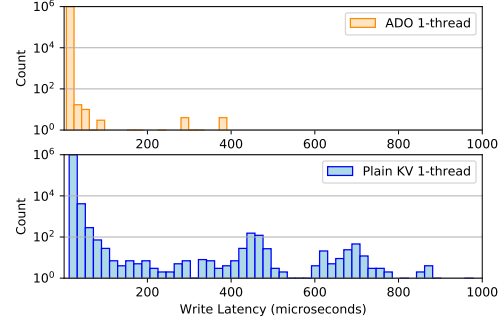


Figure 16: Comparison of Write Latency for ADO vs. Plain-KV Deployments

total number of threads, can sustain 43% and 30% higher throughput than Plain-KV for 2 threads/client and 1 thread/client respectively.

Figure 16 compares the write latency distribution for the two deployments. For the ADO deployment, 99.97% are less than  $20\mu\text{sec}$  and for Plain-KV, 94.27% are less than  $20\mu\text{sec}$ . The mean latencies are  $6.7\mu\text{sec}$  and  $16.6\mu\text{sec}$  for the ADO and Plain-KV deployments respectively. The increased latency observed by Plain-KV is due to the fact that in addition to remote write updates, the single client thread is performing local data structure updates, remote summaries updates and erases for aging-out of the data (all on the same shard).

**Query Latency** Next, we compare the query latency for the two deployments. Figure 17 compares latencies for an 8MiB quantum. For the ADO deployment latencies vary from  $4.0\text{ms}$  to  $115.0\text{ms}$  with a mean of  $54.4\text{ms}$ . Slightly better, the Plain-KV deployment shows latencies varying from  $2.8\text{ms}$  to  $103.0\text{ms}$ , with a mean of  $52.8\text{ms}$ . We consider this improvement of negligible value in practice.

**Memory Footprint** Finally, we compare the memory footprint for each deployment. Quanta are 16MiB (262K records) and they are age-out at a limit of 10. A single shard with

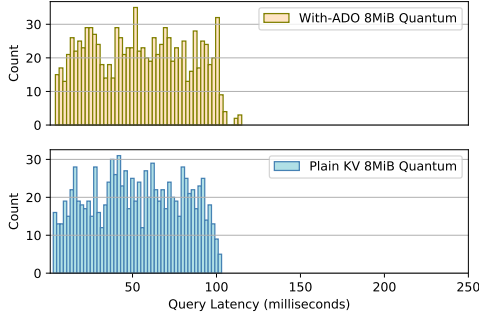


Figure 17: Comparison of Query Latency for ADO vs. Plain-KV Deployments

single client thread arrangement is used.

	ADO	Plain-KV
Client DRAM	912 MiB	1747 MiB ( $\uparrow$ 91%)
Server DRAM	530 MiB	651 MiB ( $\uparrow$ 23%)
Server Optane PM	327 MiB	655 MiB ( $\uparrow$ 100%)
Total Memory Footprint	1.72 GiB	2.98 GiB ( $\uparrow$ 72%)

Table 3: Memory Footprint Comparison

The data in Table 3 provides a comparison of client and server side memory footprints for the different schemes. Note, the seemingly large client-side footprint is due to the replay data, which is pre-loaded before execution of the benchmark (in both instances).

The ADO deployment is more memory efficient mainly because there is only a single copy of the data. However, the Optane-PMM footprint for the Plain-KV deployment is twice that of the ADO deployment. This is caused by the metadata overhead incurred by saving updates as individual entries in the key-value hash table - in the ADO deployment the quantum is a contiguous array of records in memory.

## 5 Related Work

In this section, we review related work on the topics of high performance in-memory key-value stores, use of persistent memory as the storage medium and the use of in-store operations.

**In-memory key-value stores** Stanford University’s RAMCloud [24] is a distributed key-value store that exploits DRAM for high-performance storage together with replication to provide durability. It focuses on maintaining in-memory low-latency while also achieving scalability and durability. Similarly, FaRM [12, 13], HERD [18] and MICA [22] further optimized latency by leveraging high-performance RDMA for networked in-memory key-value stores. All of these are designed for DRAM and have demonstrated sub-millisecond latencies. However, they are not able to manage crash-consistency and persistent data structures.

**Persistent memory aware data stores** Other work takes existing data stores and modifies the back-end to support Optane-PMM, which means replacing read/write I/O instructions with direct load/store memory instructions. Izraelevitz et. al [16] change the back-end of Redis, RocksDB, and MongoDB to support Optane-PMM.

The general outcome from this work is a demonstration of better performance when using persistent memory directly with DAX mode (as opposed to block mode).

More recently, PMemKV is a new Optane-PMM-optimized key-value data-store [1] developed by Intel. Its storage engine can be replaced by various tree data structures provided by PMDK [4]. Izraelevitz et. al [16] benchmarked different storage engines and measured random read latency at around  $1\mu sec$  and write latency at around  $4\mu sec$ . Note, this benchmark provides only local performance for a single thread; the system is not networked.

**Remote memory paradigm** AsymNVM [23] focuses on providing RDMA-like operations but with transactional semantics. The key idea to remove persistency bottleneck is the use of an operation log that reduces stall time due to RDMA writes and enables efficient batching and caching in front-end nodes. AsymNVM is based on a remote memory paradigm; there is no index or key space. Furthermore, AsymNVM does not allow operations close to the memory.

Clover [32] offers an exploration of passive disaggregated persistent memory. They do provide space management but their focus is primarily a remote memory paradigm in that compute is performed on the client side and memory is attached remotely.

StRoM (Smart Remote Memory) [28] investigates the use of FPGA-based computation in the smart NIC card to perform operations on remote memory. Limited by the FPGA, they provide only fundamental operations on the data (e.g., pointer-chasing, filtering, aggregation).

**In-store operations** Prior work in “shipping” operations to a data store focuses on using volatile memory combined with conventional storage. To our knowledge, MCAS is the first work to support in-store operations in persistent, byte addressable memory.

Stored procedures are widely adopted in the database community. They are typically in the form of programs defined in a programming language that are translated (as opposed to natively compiled) at runtime [5, 6]. To improve performance, stored procedures can be translated prior to execution into compact and more efficient machine code [14]. H-Store [19], VoltDB [30] and Hazelcast [27] are examples of databases that use Java-based stored procedures.

Redis [9] pioneered early work on supporting user-defined extensions in key-value stores. Redis allows customized modules with built-in data structures. The modules are able to manipulate data structures directly in the store.

However, Redis is limited to built-in data structures in contrast to MCAS’s ability to support any native pointer-based data structure.

Splinter [20] is another in-memory key-value store that supports natively executable extensions. It supports any customized operations on arbitrary data structures written in Rust [2]. However, unlike MCAS, user-defined extensions in Splinter cannot directly operate on value data structures (it does not use memory mapping). Instead, it copies data to commit changes. Splinter also uses DRAM and therefore does not need to deal with data persistency and crash-consistency.

## 6 Conclusion and Future Work

In this paper we have outlined an evolution of the traditional key-value store paradigm with near-data compute in the form of Active Data Objects. To our knowledge, this is a first-of-its-kind technology that combines RDMA networking, persistent memory and crash-consistent in-store programmability. We have evaluated our approach in the context of a real-world use case based on indexing for Continuous Data Protection (CDP). We show that the ADO-based approach results in up to 43% higher throughput performance and 72% less memory footprint compared to the Plain-KV approach. In terms of raw performance numbers, with three-way replication, our solution can over 6M updates/sec in a three node cluster of 2-socket servers.

The low read-write amplification and low latency enabled by Optane-PMM, in conjunction with RDMA, allows the system to support a synchronous “memory like” behavior that simplifies design and consistency models. With three replicas (immediately consistent), MCAS can potentially support thousands of persistent volumes with an aggregate IOPS throughput of over 6M IOPS in a 2-socket server.

A key design element of MCAS is its Active Data Object (ADO) plugin mechanism. This feature allows the system to extend the key-value paradigm with arbitrary in-place opera-

tions that can be safely executed (using process isolation) in the MCAS system. In the current prototype, users implementing ADO operations use either hand-crafted code or STL instrumented with undo log capabilities. We believe that a key hurdle for the adoption of persistent memory is making it easier to develop crash-consistent code.

From the perspective of furthering our solution, we plan to explore the use of new languages, compilers and libraries as well as hardware support for transparent undo logging [8, 11]. Extending these concepts to allow for secure (e.g. enclave-like) execution in the storage system is also a topic of further work.

## Availability

MCAS is an open source project maintained at <https://github.com/IBM/mcas/>. Additional information about the project can be gained from <https://github.com/IBM/mcas/>.

## References

- [1] Intel corporation - pmemkv. <https://github.com/pmem/pmemkv>.
- [2] Rust programming language. <https://www.rust-lang.org/>. Accessed: 2020-9-17.
- [3] 3d xpoint Optane Media Memory MarketsBits, Revenue, Costs. [http://mkwventures.com/uploads/20201110\\_SESSION\\_A-3\\_WEBB.pdf](http://mkwventures.com/uploads/20201110_SESSION_A-3_WEBB.pdf), 2020.
- [4] Intel corporation. Persistent Memory Development Kit. <http://pmem.io/pmdk/>, 2020.
- [5] Microsoft, inc. Transact-SQL reference (database engine). <http://docs.microsoft.com/en-us/sql/t-sql/language-reference>, 2020.
- [6] Oracle, Inc. Oracle Database 12c pl/sql. <http://www.oracle.com/technetwork/database/features/plsql/index.html>, 2020.
- [7] Spectra - Digital Data Outlook 2020. [https://spectralogic.com/wp-content/uploads/digital\\_data\\_storage\\_outlook\\_2020.pdf](https://spectralogic.com/wp-content/uploads/digital_data_storage_outlook_2020.pdf), 2020.
- [8] M. Cai, C. C. Coats, and J. Huang. Hoop: Efficient hardware-assisted out-of-place update for non-volatile memory. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 584–596, 2020.
- [9] Josiah L. Carlson. *Redis in Action*. Manning Publications Co., USA, 2013.
- [10] Brian Choi, Parv Saxena, Ryan Huang, and Randal Burns. Observations on porting in-memory kv stores to persistent memory, 2020.
- [11] K. Doshi, E. Giles, and P. Varman. Atomic persistence for scm with a non-intrusive backend controller. In *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 77–89, 2016.
- [12] Aleksandar Dragojević, Dushyanth Narayanan, Orion Hodson, and Miguel Castro. Farm: Fast remote memory. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation, NSDI’14*, page 401–414, USA, 2014. USENIX Association.
- [13] Aleksandar Dragojević, Dushyanth Narayanan, Edmund B. Nightingale, Matthew Renzelmann, Alex Shamis, Anirudh Badam, and Miguel Castro. No compromises: Distributed transactions with consistency, availability, and performance. In *Proceedings of the 25th Symposium on Operating Systems Principles, SOSP ’15*, page 54–70, New York, NY, USA, 2015. Association for Computing Machinery.

- [14] Craig Freedman, Erik Ismert, and Per-Åke Larson. Compilation in the microsoft SQL server hekaton engine. *IEEE Data Eng. Bull.*, 37(1):22–30, 2014.
- [15] Maurice Herlihy, Nir Shavit, and Moran Tzafrir. Hopscotch hashing. In *Distributed Computing, 22nd International Symposium, DISC 2008, Arcachon, France, September 22-24, 2008. Proceedings*, pages 350–364, 2008.
- [16] Joseph Izraelevitz, Jian Yang, Lu Zhang, Juno Kim, Xiao Liu, Amirsaman Memaripour, Yun Joon Soh, Zixuan Wang, Yi Xu, Subramanya R. Dulloor, Jishen Zhao, and Steven Swanson. Basic performance measurements of the intel optane DC persistent memory module. *CoRR*, abs/1903.05714, 2019.
- [17] Olzhas Kaiyrakhmet, Songyi Lee, Beom-seok Nam, Sam H. Noh, and Young ri Choi. Slm-db: Single-level key-value store with persistent memory. In *17th USENIX Conference on File and Storage Technologies (FAST 19)*, pages 191–205, Boston, MA, February 2019. USENIX Association.
- [18] Anuj Kalia, Michael Kaminsky, and David G. Andersen. Using rdma efficiently for key-value services. In *Proceedings of the 2014 ACM Conference on SIGCOMM, SIGCOMM ’14*, page 295–306, New York, NY, USA, 2014. Association for Computing Machinery.
- [19] Robert Kallman, Hideaki Kimura, Jonathan Natkins, Andrew Pavlo, Alexander Rasin, Stanley Zdonik, Evan P. C. Jones, Samuel Madden, Michael Stonebraker, Yang Zhang, John Hugg, and Daniel J. Abadi. H-store: A high-performance, distributed main memory transaction processing system. *Proc. VLDB Endow.*, 1(2):1496–1499, August 2008.
- [20] Chinmay Kulkarni, Sara Moore, Mazhar Naqvi, Tian Zhang, Robert Ricci, and Ryan Stutsman. Splinter: Bare-metal extensions for multi-tenant low-latency storage. In Andrea C. Arpaci-Dusseau and Geoff Voelker, editors, *OSDI*, pages 627–643. USENIX Association, 2018.
- [21] Sheng Li, Hyeontaek Lim, Victor W. Lee, Jung Ho Ahn, Anuj Kalia, Michael Kaminsky, David G. Andersen, O. Seongil, Sukhan Lee, and Pradeep Dubey. Architecting to achieve a billion requests per second throughput on a single key-value store server platform. *SIGARCH Comput. Archit. News*, 43(3S):476–488, June 2015.
- [22] Hyeontaek Lim, Dongsu Han, David G. Andersen, and Michael Kaminsky. Mica: A holistic approach to fast in-memory key-value storage. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation, NSDI’14*, page 429–444, USA, 2014. USENIX Association.
- [23] Teng Ma, Mingxing Zhang, Kang Chen, Zhuo Song, Yongwei Wu, and Xuehai Qian. Asymnmv: An efficient framework for implementing persistent data structures on asymmetric nvm architecture. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS ’20*, page 757–773, New York, NY, USA, 2020. Association for Computing Machinery.
- [24] John Ousterhout, Arjun Gopalan, Ashish Gupta, Ankita Kejriwal, Collin Lee, Behnam Montazeri, Diego Ongaro, Seo Jin Park, Henry Qin, Mendel Rosenblum, Stephen Rumble, Ryan Stutsman, and Stephen Yang. The ramcloud storage system. *ACM Trans. Comput. Syst.*, 33(3), August 2015.
- [25] Thanumalayan Sankaranarayanan Pillai, Vijay Chidambaram, Ramnatthan Alagap-



- pan, Samer Al-Kiswany, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. Crash consistency. *Queue*, 13(7):20–28, July 2015.
- [26] Jinglei Ren, Jishen Zhao, Samira Khan, Jongmoo Choi, Yongwei Wu, and Onur Mutlu. Thynvm: Enabling software-transparent crash consistency in persistent memory systems. In *Proceedings of the 48th International Symposium on Microarchitecture*, MICRO-48, page 672–685, New York, NY, USA, 2015. Association for Computing Machinery.
- [27] Haytham Salhi, Feras Odeh, Rabee Nasser, and Adel Taweel. Open source in-memory data grid systems: Benchmarking hazelcast and infinispan. In *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering*, ICPE ’17, page 163–164, New York, NY, USA, 2017. Association for Computing Machinery.
- [28] David Sidler, Zeke Wang, Monica Chiosa, Amit Kulkarni, and Gustavo Alonso. Strom: Smart remote memory. In *Proceedings of the Fifteenth European Conference on Computer Systems*, EuroSys ’20, New York, NY, USA, 2020. Association for Computing Machinery.
- [29] Storage Networking Industry Association (SNIA). Technical whitepaper: Data protection best practices. <http://bit.ly/2GR11bW>, October 2017. Accessed: 2020-01-24.
- [30] Michael Stonebraker and Ariel Weisberg. The voltdb main memory dbms. *IEEE Data Eng. Bull.*, 36:21–27, 2013.
- [31] K. Tang, W. Tong, J. Ma, and B. Liu. Dv-nvllc: Efficiently guaranteeing crash consistency in persistent memory via dynamic versioning. In *2019 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 1–8, 2019.
- [32] Shin-Yeh Tsai, Yizhou Shan, and Y. Zhang. Disaggregating persistent memory and controlling them remotely: An exploration of passive disaggregated key-value stores. In *USENIX Annual Technical Conference*, 2020.
- [33] Daniel Waddington, Clem Dickey, Luna Xu, Travis Janssen, Jantz Tran, and Doshi Kshittij. Evaluating Intel 3D-Xpoint NVDIMM Persistent Memory in the context of a Key-Value Store. In *International Symposium on Performance Analysis of Systems and Software*, ISPASS ’20. IEEE, August 2020.