ctFS: Replacing File Indexing with Hardware Memory Translation through Contiguous File Allocation for Persistent Memory

RUIBIN LI, XIANG REN, XU ZHAO, SIWEI HE, MICHAEL STUMM, and DING YUAN,
University of Toronto, Canada

Persistent byte-addressable memory (PM) is poised to become prevalent in future computer systems. PMs are significantly faster than disk storage, and accesses to PMs are governed by the **Memory Management Unit (MMU)** just as accesses with volatile RAM. These unique characteristics shift the bottleneck from I/O to operations such as block address lookup—for example, in write workloads, up to 45% of the overhead in ext4-DAX is due to building and searching extent trees to translate file offsets to addresses on persistent memory.

We propose a novel *contiguous* file system, ctFS, that eliminates most of the overhead associated with indexing structures such as extent trees in the file system. ctFS represents each file as a contiguous region of virtual memory, hence a lookup from the file offset to the address is simply an offset operation, which can be efficiently performed by the hardware MMU at a fraction of the cost of software-maintained indexes. Evaluating ctFS on real-world workloads such as LevelDB shows it outperforms ext4-DAX and SplitFS by 3.6× and 1.8×, respectively.

CCS Concepts: • **Software and its engineering** → **File systems management**; *Allocation/deallocation strategies;*

Additional Key Words and Phrases: Persistent memory, file system, page table, memory allocation, data center

ACM Reference format:
Ruibin Li, Xiang Ren, Xu Zhao, Siwei He, Michael Stumm, and Ding Yuan. 2022. ctFS: Replacing File Indexing with Hardware Memory Translation through Contiguous File Allocation for Persistent Memory. *ACM Trans. Storage* 18, 4, Article 30 (December 2022), 24 pages.
https://doi.org/10.1145/3565026

1 INTRODUCTION

The emergence of byte-addressable **persistent memory (PM)** fundamentally blurs the boundary between memory and persistent storage. Intel's Optane DC persistent memory is byte-addressable and can be integrated as a memory module. Its performance is orders of magnitude faster than traditional storage devices: the sequential read, random read, and write latencies of Intel Optane DC are 169 ns, 305 ns, and 94 ns, respectively, which are the same order of magnitude as DRAM (86 ns) [21].

This research was supported by the Canada Research Chair fund, an NSERC discovery grant, and a VMware gift.
Authors' address: R. Li, X. Ren, X. Zhao, S. He, M. Stumm, and D. Yuan, Sandford Fleming 2002E, 10 King's College Road, Toronto, ON M5S 3G4 Canada; emails: {robinlrb.li, jenny.ren}@mail.utoronto.ca, i@xuzhao.net, siwei.he@mail.utoronto.ca, stumm@eecg.toronto.edu, yuan@ece.utoronto.ca.

1553-3077/2022/12-ART30
https://doi.org/10.1145/3565026

A number of file system designs have been introduced with the aim of exploiting the characteristics of PM. For example, Linux introduced **Direct Access support (DAX)** for some of its file systems (ext4, xfs, and ext2) that eliminates the use of the page cache. Other designs bypass the kernel by mapping different file system data structures into user space to reduce the overhead of switching into the kernel [6, 7, 23, 27, 37]. SplitFS, a state-of-the-art PM file system, aggressively uses memory-mapped I/O [23] for significantly improved performance.

All of these systems use conventional tree-based index structures for translating the file offset to the device address. This index structure was first proposed by Unix in the '70s [35] when the speed of memory and persistent storage differed by several orders of magnitude. However, with the emergence of PM, this speed difference has shrunk significantly to the point of being almost negligible. This, in turn, has shifted the bottleneck from I/O to file indexing overheads.

Indeed, we show in Section 2 that this indexing overhead can be as high as 45% of the total runtime for write workloads on ext4-DAX (e.g., for Append). While memory-mapped I/O (`mmap()`) can mitigate some of the indexing overhead [10], it does not remove indexing overhead but only shifts its timing to page fault handling or `mmap()` (when pre-fault is used). For example, Section 2 shows that with SplitFS, file indexing overhead can be as high as 63% of the Append workload runtime. This is 18% higher than that of ext4-DAX, even though the runtime of Append is lower on SplitFS; this is because SplitFS's improved performance further shifts the bottleneck and exacerbates indexing overhead.

An alternative to using file indexing is to use contiguous file allocation. While simple contiguous allocation designs with fix-size or variable-size partitions are known [36], they face two major design challenges: (1) internal fragmentation for fix-size partitions, (2) external fragmentation for variable-size partitions, and (3) file resizing, specifically for expansion, which often requires costly data movement. Therefore, the only use of contiguous file allocation in practice today is on CD-ROMs, where files are read-only [36]. SCMFS [39] proposed the high-level idea of allocating files contiguously in virtual memory. However, it does not address the challenges of contiguous file allocation, namely, how files are allocated and how resizing is managed.

We propose ctFS, a contiguous file system designed from the ground up for PM. ctFS has the following key design elements:

- Each file (and directory) is contiguously allocated in the 64-bit *virtual* memory space. We demonstrate the practicality of this idea, given that the 64-bit address space is enormous. Furthermore, the virtual address space is carefully managed by a hierarchical layout, similar to the buddy memory allocation [25], in which each partition is subdivided into 8 equal-size sub-partitions. This design speeds up allocation, avoids external fragmentation, and minimizes internal fragmentation (Section 3.2).
- A file's virtual-to-physical mapping is managed using ***persistent page tables* (PPT)**. PPTs have a similar structure as the regular, volatile page tables in DRAM, except that PPTs are stored persistently on PM. Upon a page fault on an address that is within a ctFS's memory region, the OS looks up the PPT and creates the same mappings in the DRAM-based page tables. Therefore, subsequence accesses are served by hardware MMU from DRAM-based page tables, avoiding the indexing cost.
- Initially, a file is allocated within a partition whose size is just large enough for the file. When a file outgrows its partition, it is moved to a larger partition in virtual memory without copying any physical persistent memory. ctFS does this by remapping the file's physical pages to the new partition using *atomic swap*, or `pswap` (Section 3.3), a new OS system call that *atomically* swaps the virtual-to-physical mappings. Atomic swap also enables efficient crash consistency

on multi-block writes without needing to double-write the data. An atomic write in ctFS simply writes the data to a new space and then pswaps it with the old data (Section 3.4).

In ctFS, the translation from file offset to the physical address now needs to go through the virtual-to-physical memory mapping, which is no less complex than the conventional file-to-block indexes. The key difference is that page translation can be sped up by existing hardware support. Translations that are cached by TLB will be handled transparently from the software and completed in one cycle. In contrast, a file system's file-to-block translation can only be cached by software. Additionally, ctFS can adopt various optimizations for memory mapping, such as using huge pages, to further speed up a variety of operations.

Our evaluation on Intel Optane DC reveals that ctFS can eliminate most indexing overheads, which results in up to a 7.7× and 3.1× speedup over ext4-DAX and SplitFS [23] on the Append workload. ctFS further improves the throughput of LevelDB running YCSB by up to 3.62×, 1.82×, 3.21×, and 2.45× when compared to ext4-DAX, SplitFS, Nova [40], and PMFS [7], respectively. Finally, ctFS improves RocksDB [19] performance by up to 1.6× when compared to ext4-DAX. The source code of ctFS is available at https://github.com/robinlee09201/ctFS.

A limitation of ctFS is that we implement it as a user-space library file system that trades protection for performance. While this squeezes the most performance out by aggressively bypassing the kernel, it sacrifices protection in that it only protect against unintentional bugs instead of intentional attacks. We envision that this is an acceptable, or even desirable, tradeoff for data center environments. We discuss other limitations in Section 5.

2 UNDERSTANDING FILE INDEXING OVERHEAD

We analyzed the performance overhead of block address translation in Linux's ext4-DAX and in SplitFS [23]. Ext4-DAX is the port of the ext4 extent-based file system to PM. It eliminates the page cache and directly accesses PM using memory operations (`memcpy()`).

Background on SplitFS. We briefly describe SplitFS for a better understanding. SplitFS splits the file system logic into a user-space library (U-Split) and a kernel space component (K-Split), where K-Split uses ext4-DAX. A file is split into multiple 2 MB regions by U-Split, where each region is mapped to one ext4-DAX file. Both U-Split and K-Split participate in indexing: U-Split maps a logical file offset to the corresponding ext4-DAX file, and the ext4-DAX in K-Split further searches its extent index to obtain the actual physical address.

SplitFS also proposed a novel operation called `relink` to improve the performance of file expansion and provide crash consistency on file writes without double-writing data. Under its sync mode, file appends are first made to a staging file and then relinked to the target file either when `fsync()` gets called or the staging file reaches its size limit; file overwrites are applied in-place. Under its strict mode, every file write, whether it is overwriting or appending data, is applied to a staging file and gets relinked at the end of every write. Hence, the indexing time of SplitFS consists of `relink`, mmap, and indexing in both kernel and user components.

Experimental Methodology. Our experiments were conducted on a server with two 128 GB Intel Optane DC **persistent memory (PM)** modules, an 8-core Intel Xeon 4215 CPU running at 2.5 GHz, and 96 GB of DRAM. We used Linux version v5.7.0-rc7+.

We ran a total of six tests. The results are shown in Figure 1. Each test either reads or writes a 10 GB file. The first test, Append, repeatedly *appends* 4 KB of data to a file that is initially empty. The second test, SWE, sequentially writes a total of 10 GB of data to an *empty* file with 10 `pwrite()` calls to write 1 GB at a time. RR reads 4 KB of data from a random (4 KB-aligned) offset in a 10 GB file, and RW overwrites an existing 10 GB file with 4 KB of data at a random (4 KB-aligned) offset,

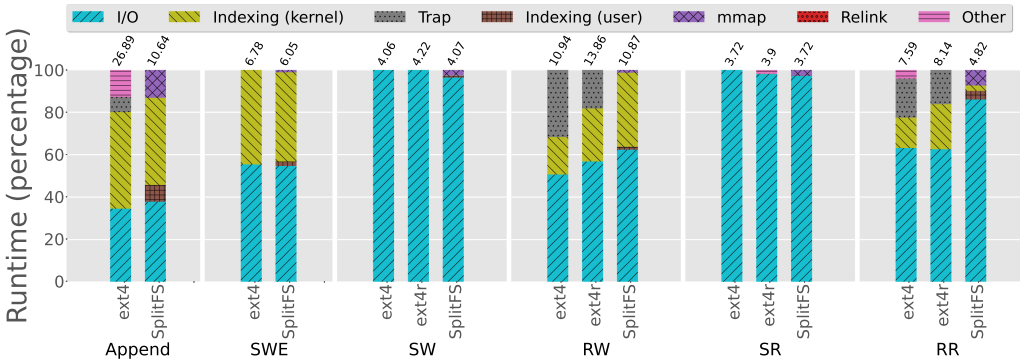

Fig. 1. Performance breakdown (in percentage) of ext4-DAX and SplitFS on persistent memory. The number above each bar is the total runtime in seconds.

and they do this 2,621,440 times. Finally, SR/SW are sequentially reads/writes 10 GB data, 1 GB at a time.[1]

For the SW, RW, RR, and SR tests, we ran the ext4-DAX tests with two types of files: those that were *sequentially* allocated (ext4) and those that were *randomly* allocated (ext4r). Sequentially allocated files were created by SWE, which maximizes ext4-DAX's grouping of blocks into an extent. Randomly allocated files were created by writing to them similarly to the way RW does, except that the file is initially empty (Linux file systems support sparse files); these randomly allocated files limit ext4-DAX's ability to group blocks into extents. The "ext4r" bars in RW, RR, and SR represent tests that operated on such randomly allocated files. Note that ext4-DAX creates 12 extents for a sequentially allocated 10 GB file, but creates 256 extents for a randomly allocated file. For SplitFS, all files are sequentially allocated.

Indexing overhead in ext4-DAX. Figure 1 shows the breakdown of the completion time of each test. For ext4-DAX, we observe that indexing overhead is significant in Append and SWE, spending at least 45% of the total runtime on indexing.[2]

For the random access workloads, RR and RW, the proportion of time spent on indexing is lower, but still considerable: 25% and 21% of the total runtime when randomly writing and reading to/from a randomly allocated file (ext4r), and 18% and 15% when the file was sequentially allocated.

Indexing Overhead in SplitFS. Figure 1 also shows the breakdown of the completion time of SplitFS's sync mode.[3] Compared to ext4-DAX, SplitFS spends an even higher proportion of the total runtime on indexing in the Append (63%), SWE (45%), and RW workloads (38%), while it spends 14% of the runtime on indexing in RR.

To understand SplitFS's indexing overhead in more detail, consider the Append workload where SplitFS spends a total of 6.62s on indexing. Three components make up this file indexing time: (1) the kernel indexing time as part of page fault handling (4.37 s), (2) U-split's file indexing time (0.84 s) spent on mapping file offsets to the correct ext4-DAX file, and (3) U-Split's `mmap()` time (1.39 s).

[1]We found that the version of SplitFS we tested does not support append operations that write over 128 MB under its sync mode. Therefore, in SWE, we write 128 MB at a time in SplitFS, instead of 1 GB as in ext4-DAX and other the file systems we discuss in Section 4.

[2]In both cases, the index time includes the time to build the index.

[3]We only show its sync mode result, as its semantics is comparable to that of ext4-DAX. SplitFS's strict mode is further evaluated in Section 4.

Table 1. The Two Modes Provided by ctFS

Mode	Atomicity		Similar to
	data	metadata	
sync	✗	✓	NOVA-relaxed, PMFS, SplitFS-sync
strict	✓	✓	NOVA-strict, Strata, SplitFS-strict

3 DESIGN & IMPLEMENTATION OF CTFS

This section starts with an overview of ctFS. Then, we describe the file system layout (Section 3.2) and how ctFS interacts with the kernel's memory management system (Section 3.3). We then explain ctFS's primitive for atomic operations—`pswap()` and how ctFS handles file updates and ensures crash consistency (Section 3.4). Finally, we discuss some optimizations (Section 3.5) and the protection model (Section 3.6).

3.1 Design Overview

ctFS is a high-performance PM file system that directly accesses and manages both file data and metadata in user space. Each file is stored contiguously in virtual memory, and ctFS offloads traditional file systems' offset to block number indexing to the memory management subsystem. ctFS achieves the following design goals:

- **POSIX compliance:** ctFS currently supports over 30 commonly used functions from the POSIX-compatible file system API.
- **Synchronous writes:** Write operations on ctFS are always synchronous, i.e., writes are persisted on PM before the operation completes. Hence, there is no need for `fsync` (which does nothing in ctFS).
- **Crash consistency:** ctFS supports both file data consistency (by using `pswap`) and metadata consistency (by using conventional redo logs).
- **Concurrent operations:** ctFS supports concurrent operations on different files or concurrent reads on the same file; a reader-writer lock is used for each file to synchronize concurrent accesses.

Similar to prior systems, such as NOVA [40] and SplitFS [23], ctFS offers two modes, sync and strict, as shown in Table 1. Both modes ensure atomic metadata operations that include directory operations. Strict mode further ensures file data writes are atomic (by using `pswap`).

ctFS's architecture, shown in Figure 2, consists of two components: (1) the user space file system library, ctU, which provides the file system abstraction, and (2) the kernel subsystem, ctK, which provides the virtual memory abstraction. ctU implements the file system structure and maps it into the *virtual* memory space. ctK maps virtual addresses to PM's physical addresses using a ***persistent page table (PPT)***, which is stored in PM. Any page fault on a virtual address inside ctU's address range is handled by ctK. If the PPT does not contain a mapping for the fault address, then ctK will allocate a PM page, establish the mapping in the PPT, and then copy the mapping from the PPT to the kernel's DRAM page table, allowing virtual to PM address translations to be carried out by the MMU. When any mapping in the PPT becomes obsolete, ctK will remove the corresponding mapping from the DRAM page table and shoot down the mapping in the TLBs.

With this architecture, there is a clear separation of concerns. ctK is *not* aware of any file system semantics, which is entirely implemented by ctU using memory operations. Next, we discuss the designs that are specific to ctFS. We omit the designs that are similar to existing file systems. For

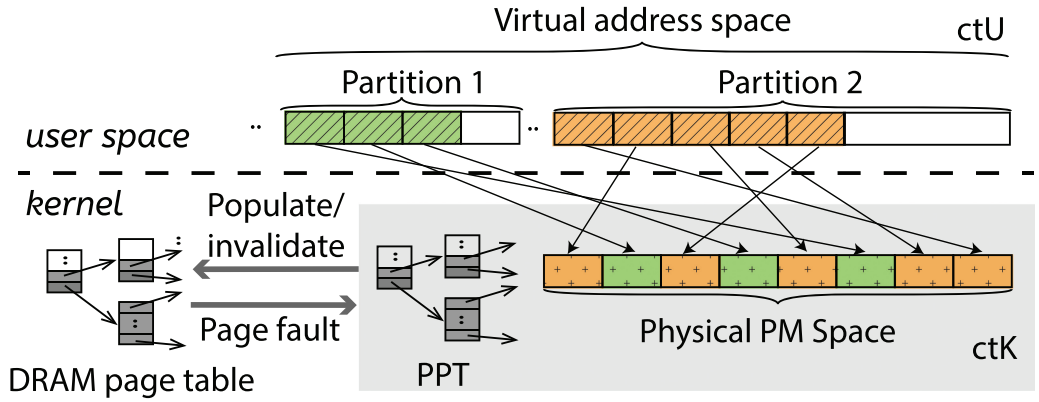

Fig. 2. Architecture of ctFS. Each box represents a page. Two partitions are shown. The file allocated in partition 1 uses 3 pages (green), and the file in partition 2 uses 5 pages. ctK maintains virtual-to-physical page mappings in the PPT.

example, we use standard transaction logging to provide crash consistency of *metadata*, including directories, inode, and ctFS data structures such as partition headers, bitmaps, and so on.

3.2 File System Structure (ctU)

ctFS's user-space library, *ctU*, organizes the file system's *virtual* memory space into hierarchical partitions to facilitate contiguous allocations. The size of each partition at a particular level is identical, and each level's size is 8× the size of the partitions at the next lower level. Figure 3 shows the sizes of the 10 levels that ctFS currently supports. The lowest level, L0, has 4 KB partitions, whereas the highest level, L9, has 512 GB partitions. ctFS can be easily extended to support more partition levels, e.g., L10 (4 TB), L11 (32 TB), and so on.

A file or directory is always allocated contiguously in one and only one partition, with the size of the partition being the smallest capable of containing the file. For example, a 1 KB file is allocated in an L0 partition (4 KB); a 2 GB file is allocated in an L7 partition (8 GB).

We chose each next level to be 8× the size of the previous level, because the boundary of the levels should align with the boundary of Linux page table levels (Figure 3). This enables the optimization during `pswap` we describe in Section 3.3. Therefore, our only options for partition size differences were: 2× (2^1), 8× (2^3), or 512× (2^9). We chose 8×, because 2× would be too small and 512× too large.

File System Layout. Figure 4 shows the layout of ctFS. The virtual memory region is partitioned into two L9 partitions. The first L9 partition is a special partition used to store file system metadata: a superblock, a bitmap for inodes, and the inodes themselves. Each inode stores the file's metadata (e.g., owner, group, protection, size) and *a single field* identifying the virtual memory address of the partition that contains the file's data. The inode bitmap is used to track whether an inode is allocated or not. The second L9 partition is used for data storage.[4]

Each partition can be in one of the three states: **Allocated (A)**, **Partitioned (P)**, or **Empty (E)**. A partition in state A is allocated to a single file; a partition in state P is divided into eight next-level partitions. We call the higher-level partition the *parent* of its eight next-level partitions. This parent partition *subsumes* its eight child partitions; i.e., these eight child partitions are sub-regions within the virtual memory space allocated to the parent. For example, in Figure 4, an L9 partition

[4]Note that the 512 GB allocated for metadata is virtual memory; the physical pages underneath it are allocated on demand.

Fig. 3. Size of partitions at levels L0 to L9. PGD, PUD, PMD, and PTE refer to the four levels of page tables in Linux (from highest to lowest). An L9 partition aligns with PGD, i.e., its starting address has zero in all of the lower-level page tables (PUD, PMD, PTE); Similarly, L6–L8 partitions align with PUD, whereas L3–L5 partitions align with PMD.

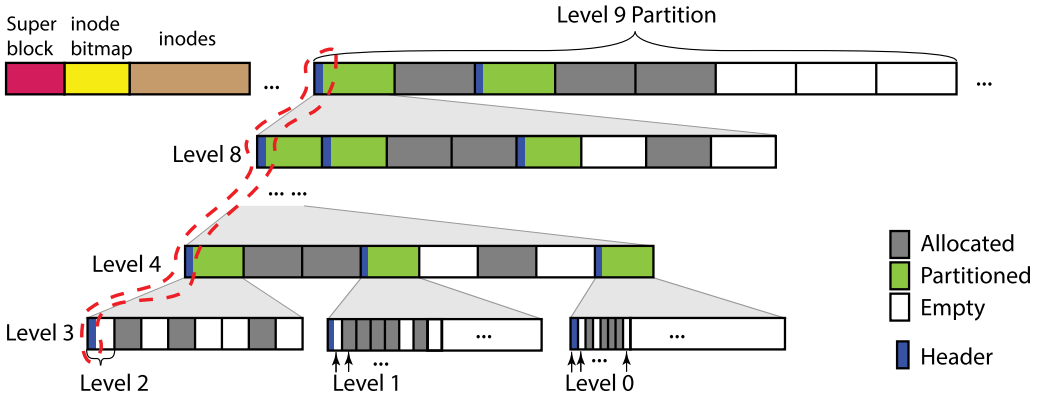

Fig. 4. Layout of ctFS in the *virtual* address space (VAS). The space of an entire partition is reserved in VAS, whereas the physical PM space is allocated on-demand based on actual usage. Headers circled in the dashed-line reside on the same page.

in state P is divided into eight L8 partitions. The first L8 partition is also in state P, which means it is divided into eight L7 partitions, and so on. In this manner, the different levels of partitions form a hierarchy.

This hierarchy of partitions has three properties. (1) For any partition, all of its ancestors must be in state P; and any partition in the A or E state does not have any descendants. (2) Any address in a partition is also an address in the partitions of its ancestors; e.g., any L3 partition in Figure 4 is contained in its ancestor L4–L9 partitions. (3) The starting address of any partition, regardless of its level, is aligned to its partition size; this is the case as long as the top-level L9 partitions are 512 GB aligned.

Partition Headers. ctU needs to maintain book keeping information for each partition, such as its state. To store such metadata, each partition in P-state has a header that contains the state of each of its *child* partitions; ctU stores the header directly on the first page of the partition for fast lookup that does not involve indirections. For example, for each partition in P state at levels L4–L9, the state of its eight children are encoded using 2 bits packed into a `uint16_t`. For an L3 partition, it uses a maximum of 64 bytes (512 bits), since it can have at most 512 children, and only 1 bit is needed for the state of each child (as it can only be A or E, but not P).

To speed up allocation, the header also has an availability level field that identifies the highest level at which a *descendent* partition is available for allocation. For example, the availability level of the left-most L9 partition in Figure 4 is 8 because this L9 partition has at least one L8 child partition in E state. With this information, when allocating a level-N partition, if a P partition's availability level is less than N, then ctU does not need to drill down further to check its child

partitions. This results in constant worst-case time complexity for allocating a parition and is far more efficient than using bitmaps.

Because ctU places the header in the first page of a partition in P state, its first child partition will also contain the same header, and as a result, this first child partition must also be in P state; it cannot be in the Allocated state, because the first page would need to be used for file content. Therefore, a header page can contain the headers of multiple partitions in the hierarchy. For example, in Figure 4, the headers in the dashed circle are all stored on the same page. This is achieved by partitioning the header page into non-overlapping header spaces for each level from L4–L9.

ctU does not partition L0–L3 further, as the 4 KB header space becomes much more wasteful for smaller partition sizes. Instead, each L3 partition (2 MB) can only be partitioned as (1) 512 L0 child partitions, (2) 64 L1 child partitions, or (3) 8 L2 child paritions, as shown at the bottom of Figure 4. As a result, there is only one header in each L3 partition that is in state P, and it contains a bitmap to indicate the status of each of its child partitions, which can only be in either state A or E, but not P.

Virtual Memory Allocation. During system initialization, ctU allocates a 1 TB, empty (i.e., not backed) **virtual memory area (VMA)** to accommodate two L9 partitions. It does not restrict the starting address of this VMA, so it can be anywhere in the virtual address space (as long as it is aligned). If the PM size is larger than 512 GB, then the next level (L10) would be used and an 8TB VMA would be allocated. Note that subsequent virtual memory allocations made from the kernel or processes will not clash with ctU's VMA, because the Linux kernel's VMA allocation will only allocate a VMA if it does not conflict with existing VMAs.

TLB usage. ctFS does not use more TLB entries compared to other file systems. In conventional (non-DAX) file systems, the file data will be buffered in memory, either in the file system's buffer cache or by the process in the case of memory mapped I/O. Such buffering will occupy TLB entries just as ctFS does, and the number of entries used depend on the amount of data a process accesses. Ext4-DAX eliminates the buffer cache by directly accessing the devices using statically mapped virtual kernel addresses. However, this mapping still goes through the page table [13] and hence still occupies TLB entries. Therefore, even compared to ext4-DAX, ctFS does not use more TLB entries.

3.3 Kernel Subsystem Structure (ctK)

ctK manages the PPT. PPT is essentially identical to a regular Linux 4-level DRAM page table, except (1) it is persistent and (2) it uses relative addresses for both virtual and physical addresses. It uses relative addresses, because ctFS's memory region may be mapped to different starting virtual addresses in different processes due to *Address Space Layout Randomization* [5, 8], and hardware reconfiguration could change PM's starting physical address. We also note that whereas each process has its own DRAM page table, ctK has a single PPT that contains the mapping of all virtual addresses in ctU's memory range (i.e., those inside the partitions). The PPT cannot be accessed by the MMU, so mappings in the PPT are used to populate entries in the DRAM page table on demand as part of page fault handling.

Note that both the allocation and populating the DRAM page table will always occur at 2 MB granularity. Whenever ctFS needs to allocate a 4 KB page, it allocates an aligned 2 MB chunk. This results in adding a new PMD entry into the PPT together with allocating a new page for the last level page table. Similarly, whenever ctFS populates the DRAM page table, it will populate the mapping of 512 base pages that are mapped by one PMD entry.

3.3.1 pswap(). ctK provides a `pswap` system call that atomically swaps the mapping of two same-sized contiguous sequences of virtual pages in the *PPT*. It has the following interface:

```
int pswap(void* A, void* B, unsigned int N, int* flag);
```

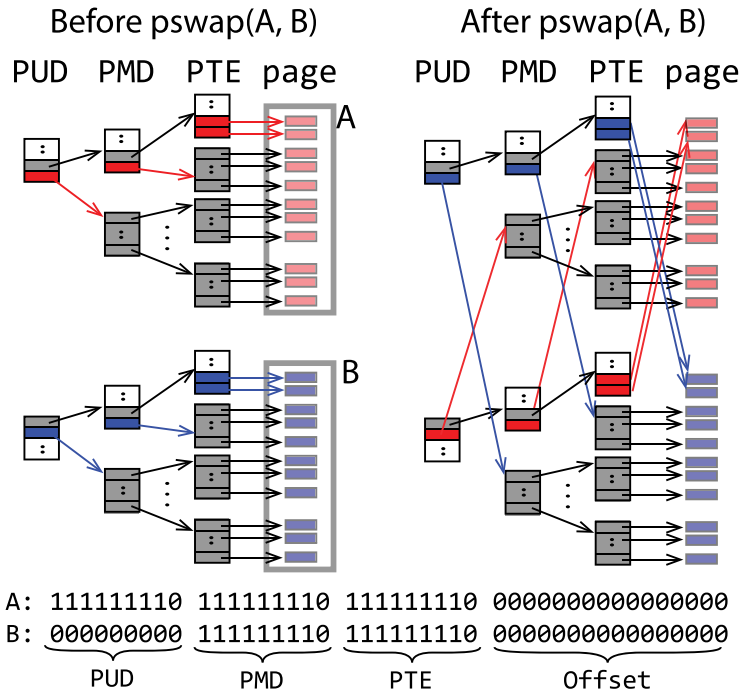

Fig. 5. An example of pswap. The shaded entries in the page tables are the ones used to map the two-page arrays A and B. The red and blue page table entries are the ones that are modified by pswap. Before pswap, A maps to the red pages and B maps to the blue pages, whereas after pswap A maps to blue pages and B maps to red pages. The last 39 bits of A and B's address are shown at the bottom.

A and B are the starting addresses of each page sequence, and N is the number of pages in the two sequences. The last parameter `flag` is an output parameter. Regardless of its prior value, `pswap` will set `*flag` to 1 if and only if the mappings are swapped successfully. ctU sets `flag` to point to a variable in the redo log stored on PM and uses it to decide whether it needs to redo the pswap upon crash recovery. `pswap` also invalidates all related DRAM page table mappings (and shoots them down in TLBs), as we found it is more efficient than updating the mappings.

The `pswap()` system call *guarantees crash consistency*: It is atomic, and its result is durable as it operates on PPT. Moreover, concurrent `pswap()` operations occur as if they are serialized, which *guarantees isolation* between multiple threads and processes.[5]

`pswap` avoids swapping every target entry in the PTEs (the last level page table) of the PPT whenever possible. Figure 5 shows an example where `pswap` needs to swap two sequences of pages—A and B—each containing 262,658 ($512 \times 512 + 512 + 2$) pages. `pswap` only needs to swap four pairs of page table entries or directories (as shown in red and blue colors in Figure 5), as all 262,658 pages are covered by a single PUD entry (covering 512×512 pages), a single PMD entry (covering 512 pages), and two PTE entries (covering 2 pages).

`pswap()` can only perform this optimization if the starting addresses of the two page sequences are *swap-aligned*. We first define the *reach* of a page table *level* to be the size of the memory region that each entry maps—e.g., the reach of PTE, PMD, PUD, and PGD are 4 K (bytes), 2 M, 1 G, and 512 G, respectively. Given two contiguous sequences of pages in virtual memory that start at

[5]`pswap` uses conventional redo log to ensure crash consistency.

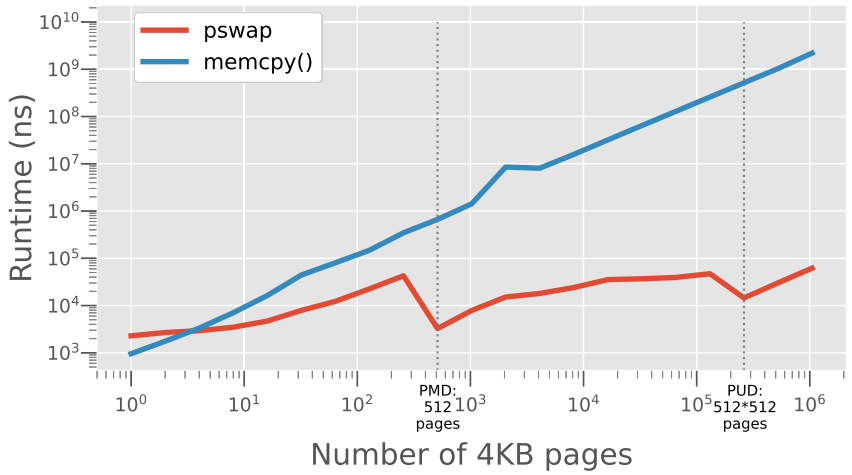

Fig. 6. Comparing the performance of pswap and memcpy. Both the X and Y axes are log scale.

addresses A and B, and given that each sequence spans a memory region of size S, let L be the highest level in the page table such that $reach(L) \leq S$. We then say that the two page sequences A and B are swap-aligned if and only if:

$$A \bmod reach(L) = B \bmod reach(L).$$

In the example of Figure 5, L is PUD, and $reach(L)$ is 1G (2^{30}). $A \bmod reach(L)$ equals $B \bmod reach(L)$, because the last 30 bits of A and B are the same.

Figure 6 shows the performance of pswap as a function of the number of pages that are swapped. We compare it with the performance of the same swap implemented with memcpy that approximates the use of conventional write ahead or redo logging that requires copying *data* twice. The pswap curve shows a wave-like pattern: As the number of pages increases, pswap latency first increases and then drops back as soon as it can swap one entry in a higher-level page table instead of 512 entries in the lower-level table. The two drop points in Figure 6 are when N is 512 (mapped by a single PMD entry) and 262,144 (mapped by a single PUD entry). In comparison, memcpy's latency increases linearly with the number of pages. When N is 1,048,576 (representing 4 GB of memory), memcpy takes 2.2 seconds, whereas pswap takes only 62 μs. However, when N is less than 4, memcpy is more efficient than pswap.

pswap() uses a redo log to ensure crash consistency. It first writes the affected page mappings to the log and then applies the changes and releases the log. If the crash happens before logging has completed, then everything remains unchanged. If it happens after the log is written, ctFS will apply the new changes upon recovery.

Concurrent invocations to pswap() will only be serialized if they operate on overlapping memory ranges. We use a binary search tree to store the ranges of concurrent, on-going pswap()s.

3.4 File System Operations

Since files are contiguous in virtual memory, read and write operations require special treatment. Other operations that operate on metadata (i.e., directories and metadata in inodes) are similar to those on conventional file systems.

Figure 7 shows how `read()` is implemented in ctFS. Given the file `offset` (from the file descriptor), ctU locates the inode and further locates the starting address of the file. It adds `offset`

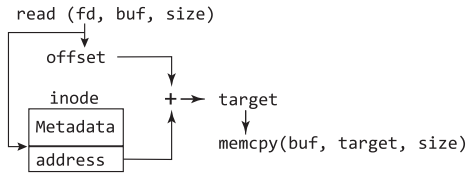

Fig. 7. Implementing `read()` on ctFS.

to this starting address, which is the virtual address of the data to be read. Then, it uses a single `memcpy()` to copy the data to the user buffer. *All of these operations are performed by the user space ctU.*

ctFS allocates a partition on-demand on the first write to a file. It always allocates the smallest partition that is big enough to store the write. Later when the file size increases beyond the partition size, ctFS will "upgrade" it to the next higher-level partition that can accommodate the file. Also recall that ctFS supports two modes, where strict mode offers atomic writes. Consequently, there are two special write cases: one that triggers an upgrade and one that requires atomicity. In the normal case where neither applies, write does not differ from read. Both of the special cases use `pswap`, and in both cases ctU *guarantees that the starting addresses are swap-aligned.* Next, we explain each case.

3.4.1 Write with Upgrade. When a write (append) triggers an upgrade, ctFS will first relocate the file to a new partition before applying the write. It also maintains a redo log to ensure crash consistency of the upgrade. Say, a write requires upgrading from a level L partition, P0, to a level M partition, P1 (where $M > L$). ctU first allocates P1 in *virtual memory*. It then calls `pswap (P0, P1, N, flag)`, where N is the number of pages in the level L partition. Note that right after the partition allocation, P1 does not map to any PM pages; therefore, after `pswap()`, P1 points to the PM pages that used to map to P0, and P0 is no longer mapped. Both steps will first be recorded in the redo log, and `flag` is located in the redo log, so if a crash occurs ctU knows whether pswap had completed successfully or not. If a crash happens before `pswap` completes, then ctU only needs to free P1. If a crash happens right after `pswap` has completed, then ctU will continue to finish the upgrade by changing the starting address in the file's inode to P1. Partition "downgrades" (e.g., when the file is truncated) are handled in a similar manner.

3.4.2 Atomic Write. In strict mode, ctFS handles an atomic write using a write-and-swap protocol. Assume a write that writes N bytes to offset O of a file in a level L partition, P0. Figure 8 shows an example, where O is *not* page aligned, and N spans three pages where the last page, p3, has not been accessed and is hence not mapped to PM. ctU carries out the following two steps:

Step 1: ctU first allocates a *staging* partition, P1, that is also at level L. It then writes the new data to the *same offset* O in P1. If O is not page-aligned, as is the case in Figure 8, then ctU copies the data fragment between the start of the first page and O in P0 to P1, and similarly, it will copy any fragment data at the end if O+N is not page aligned. Note that ctU does *not* need to copy any pages that are not modified.

Step 2: ctU invokes `pswap()` to atomically swap the page sequence in P0 with its corresponding sequence in P1. In Figure 8, it `pswaps` pages p1–p3 in partition P0 with pages p5–p7 in partition P1.

To handle crash consistency, ctU uses the redo log that records the status of each step, and the flag used in `pswap()` is located on this redo log.

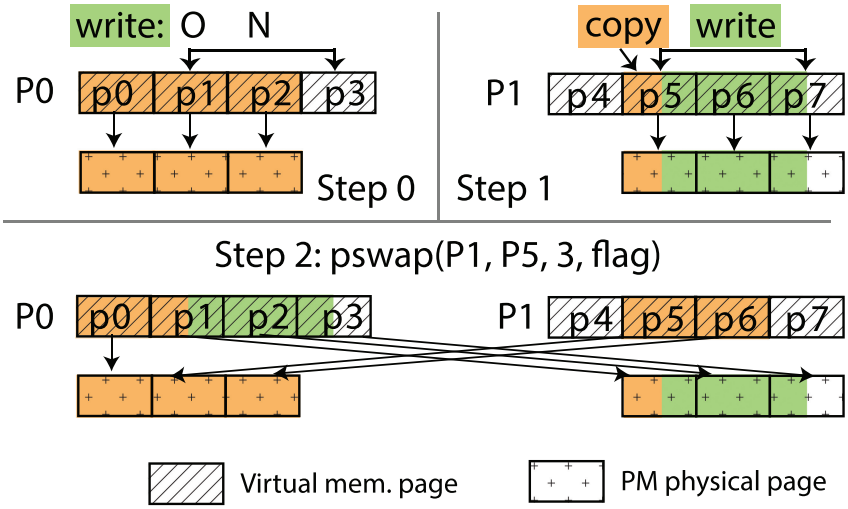

Fig. 8. An example of *atomic write* using pswap. The yellow color indicates the original file content, whereas green indicates new data to be written.

3.5 Other Optimizations

Huge Page. ctK allocates huge pages (2 MB pages) whenever possible. Because the address of any partition is aligned with the partition size, all files that reside in level L3 or above benefit from huge pages. However, when ctU performs `pswap` with small N, huge pages may have to be broken into base pages, adding extra overhead to `pswap()`. Note that `pswap` can apply its optimization whenever the page sequences are swap-aligned regardless of whether they are huge pages or not. Huge pages are enabled in our evaluation unless otherwise noted. In Section 4.1.3, we evaluate and explain the impact of huge pages in detail.

Atomic write with undo log. Figure 6 shows that `memcpy` is faster than `pswap` when the number of pages is less than 4. We further count in the impact of the write amplification and huge page breakdown and decide that we only use the write-and-swap protocol to perform atomic overwrite when the number of pages is greater than 8. Otherwise, we use undo log to preserve the original data in the staging partition first and then overwrite the file directly. This optimization is *not enabled* in the evaluation.

Optimized append in strict mode. ctFS performs an optimization on append operations by exploring the invariant between a file's metadata and its data [3, 11]. Instead of using the write-and-swap protocol, it directly appends the new data and then changes the file size in the inode. If a crash occurs before the append completes, then the file will be consistent, as the file size still has the old value, presenting a view as if the append did not occurr.

Instruction choices in `memcpy()`. We experimented with different memory copy strategies (e.g., repeat instructions, non-temporal instructions, cache flush) and found that AVX512 [20] non-temporal 512-bit load and store (VMOVDQU and MOVNTDQ) performed the best, resulting in a 5%–20% performance gain over what SplitFS and ext4-DAX uses.

Relaxed `pswap` crash consistency. Since both use cases of `pswap` (partition upgrade and atomic write) do not require the data in `A` to survive in a crash, `pswap` can eliminate the redo log. To illustrate, say we want to swap two variables: A and B. If losing data of A is acceptable, then we can simply copy A to a temporary variable T in DRAM, then assign B to A, and finally, assign T to B. In case of a crash, despite losing T (i.e., the data of original A), we can still redo assigning B to A.

This may end up in space leakage, however, data corruption is still well prevented. We consider it to be acceptable, because crashes happen rarely. However, this optimization cannot apply if pswap is used in other future features of ctFS (e.g., transaction roll-back).

DRAM page table handling during pswap. The decision whether to update Linux's DRAM page table during pswap with the modified PPT entries depends on the size of the region being swapped and the usage pattern after swapping. If the swapped region is small and accessed immediately after swapping, then updating the DRAM page tables is preferable. Otherwise, invalidating the corresponding DRAM page table entries is more efficient. Among all the workloads we encountered so far, the former case is rare. Therefore, we always invalidate the DRAM page table entries (and shoot them down in TLBs) in pswap.

Pre-populating the page table. Another optimization is to pre-populate the mappings in the DRAM page table. ctK provides a system call for pre-populating the DRAM page table mapping of a target virtual memory region. If the region is not allocated with PM pages, it will also allocate the necessary PM pages. This avoids future page faults. For large `read` and `write` operations, ctU will invoke ctK to pre-populate the mappings. In our evaluation, however, the performance difference between pre-populating and on-demand paging has shown to be negligible.

3.5.1 Optimized pswap *Usage.* While pswap is efficient when swapping a large number of pages (Figure 6), it is relatively inefficient when the number of pages is small. In addition, pswap on a small number of base pages may break up a huge page. Two optimizations can be used to avoid the pswap overhead.[6]

Lazy pswap. We can delay pswap until the time the affected pages are read, similar to the idea of *shadow paging* [30]. After new data is written to the staging partition (Step 1 in the write-and-swap protocol), we do not need to perform a pswap right away. Instead, we can mark the affected pages in the original partition (Partition 0) as invalid. This can be efficiently implemented with a bitmap using one bit for each page of a partition that tracks whether the mapping is valid. We only perform the pswap when the affected page is later read, or simply offload it to a background thread. This offers two advantages: First it avoids the overhead of pswap on the critical path; and second, it provides opportunities to batch small pswaps into a large one.

Care needs to be taken if the page is written again. In this case, we can alternate the writes between the staging partition and the original partition: The first write is to staging partition, and the subsequent write is performed on the original partition (and change the valid bit in the bitmap back to valid), and write to staging partition again.

Replace small pswap with memcpy. As shown in Figure 6, a memcpy-based swap can be more efficient than pswap when the number of pages is small. Hence, we can provide pswap as a user-space library function that chooses whether to use memcpy or invoke the kernel pswap based on the number of pages involved.

3.5.2 Bufferless Read, Write, and mmap. ctFS provides no-copy file access operations that directly return the memory address of the file. Specifically, it provides non-POSIX compatible variants of `read()` and `write()`. For `read()`, the memory address of the offset in the partition that contains the file will be directly returned to the user. In the case of `write()`, the buffer provided by the user that contains the new data will be directly mapped to the file instead of copying, using the write-and-swap protocol. (Pswap will function correctly, regardless of whether the buffer is swap-aligned or not; a swap-aligned buffer simply enables optimizations.)

For mmap, ctFS will directly map the partition to the virtual memory address. Recall that `addr`, the first argument of mmap, specifies the starting address of the new mapping. When `addr` is `NULL`,

[6] At the time of submission, we have not finished implementing these optimizations.

any address can be chosen at which to create the mapping. Hence ctFS handles `mmap` differently based on whether `addr` is NULL or not. If it is NULL, then ctU will return the partition's virtual address without invoking ctK. However, if `addr` is not `NULL`, then ctK will be invoked to map the PM pages of the file to `addr`, the same way the kernel currently handles `mmap`.

3.6 Protection

For protection, ctFS's exloits both Intel **Memory Protection Keys (MPK)** and regular page table protection. We first explain Intel MPK before discussing ctFS's design.

Background on Intel MPK. MPK allows each memory page to be tagged with one of 16 protection keys, K0, K1, . . . , K15. Four unused bits in each page table entry are used to store the key for the page. Each key's protection rights can be changed from user space, *using a special instruction*. For example, key K0's right can be set to no access, K1 can be set to read only, and K2 can be set to read/write. The access rights associated with each key are stored in a register called `PKRU`. Hence, the access rights are thread-local (as each core has its own `PKRU` register).

A key advantage of using MPK over the conventional `mprotect()` system call is performance. While assigning a key to a page still requires a system call, setting/changing the access permission associated with each key is a user-space instruction that only consumes around 20 cycles [34].

Protection in ctFS. ctFS tags each page within ctFS's memory region with a single MPK key, which we refer to as NONE. When a ctFS operation is invoked, ctU sets the access right for the NONE key to be read/write, and it resets the access right back to no access before returning. Therefore, any access to ctFS's memory space from outside of ctFS will be prevented. If multiple processes with different access rights access the same file concurrently, then ctFS can protect the same page differently for different processes as the access rights for the same key, NONE, can be set differently on different cores.

This protection strategy protects ctFS against *unintentional* bugs. For example, a dangling pointer in an application will not be able to accidentally corrupt the file system, given that changing the rights associated with the key requires a special instruction. However, this design does not protect against *intentional* attacks. For example, a malicious application could intentionally set the rights for the NONE key to be read/write and modify the file system in an arbitrary manner.

We can also extend the PPT to further include file system protection information for each page, including the access rights for owner, group, others, and the user ID and group ID. (Recall that the PPT is solely managed by the kernel subsystem and it is not visible to MMU, hence, we have the flexibility to design any structure.) We also need to add the protection checking logic into user space ctFS into the kernel subsystem. Therefore, ctFS's kernel subsystem will reject any page faults that are triggered by illegal accesses. At the time of the submission, we have not finished the implementation.

4 EVALUATION

In this section, we present the results of evaluating ctFS against other PM **file systems (FS)** using both real-world applications and microbenchmarks. The server and OS settings used in our evaluation are as described in Section 2.

In addition to comparing performance and showing that ctFS successfully removes the file indexing overhead, we also strive to answer two interrelated questions:

- How does ctFS compare with memory-mapped I/O?
- Is it fair to compare ctFS, a user-space file system, with other kernel-based file systems?

We answer the first question by carefully comparing ctFS with SplitFS [23], a state-of-the-art PM file system that aggressively uses memory-mapped I/O and kernel by-passing. For example, SplitFS

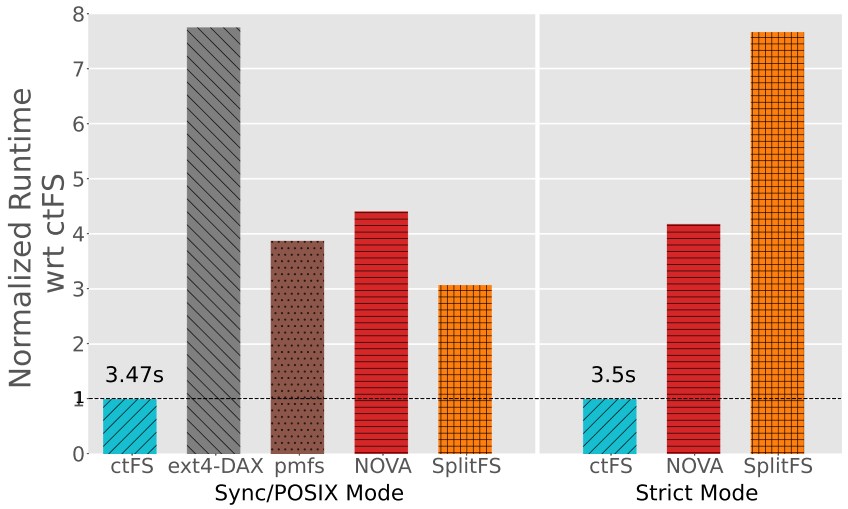

Fig. 9. Runtime of Append normalized to the runtime of ctFS. The different file system and configuration combinations are grouped by the crash consistency guarantees on file data.

mmaps the existing files and unused staging files at file system bootup time, and it uses memory accesses for `read()` and `write()` operations. Nevertheless, we note that memory-mapped I/O does *not* remove the need for file indexing, but only shifts its timing to either page fault handling time or `mmap()` time (when prefault is used). In Section 2, we showed that such file indexing causes much overhead. Similar to ctFS, SplitFS also uses 2 MB huge pages whenever possible.

We answer the second question with several efforts. First, we provide a detailed breakdown of the runtime as in Section 2, so readers can clearly see the contribution and gravity of indexing time, which is the overhead that ctFS is designed to remove. Second, we clearly show the kernel trap time, which is the "unfair" component, as ctFS performs equivalent file system operations in user-space without kernel trapping. In addition, SplitFS [23] also features aggressive kernel by-passing: only metadata modifications are routed to the kernel. Yet there are no metadata modifications in our benchmarks in Section 4.1.

4.1 Micro-benchmarks

We evaluate the performance of ctFS and compare it with that of SplitFS, ext4-DAX, PMFS, and NOVA, using the same six micro-benchmarks as in Section 2. We repeat each experiment 10 times and report the average. In all experiments, ctFS uses demand paging and does not pre-populate any pages to accentuate the maximum page fault handling overhead in ctFS. SplitFS prefaults staging files at its system bootup time so there are no page faults on those files.

4.1.1 Append. Append is particularly relevant, as the Append operation is the dominant file system operation of many application [23], and it is the operation on which SplitFS shows the most significant speedup. Figure 9 shows the performance of Append.

ctFS achieves a 7.7× speedup against ext4-DAX for Append in sync mode. 45% of ext4-DAX's runtime is in building and searching indices, as it has to allocate many small extents. Even if we deduct kernel trap overhead (shown in Figure 1) from the runtimes of ext4, ctFS-sync still achieves a 7.0× speedup. This shows the benefit of using contiguous file allocation, regardless of whether it is a user-space or kernel-space implementation.

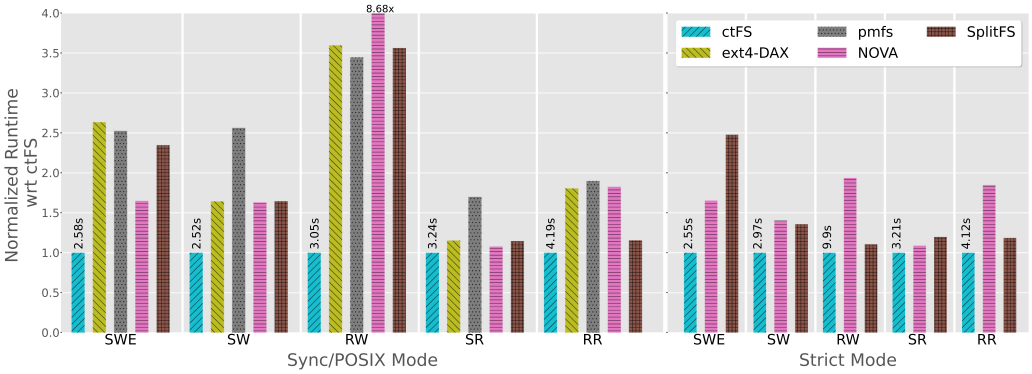

Fig. 10. Performance comparison of ext4-DAX, Nova, PMFS, SplitFS, and ctFS for five micro-benchmarks.

While SplitFS is able to significantly reduce the indexing time by using memory-mapped I/O, ctFS still achieves 3.1× speedup over SplitFS in sync mode. Specifically, SplitFS takes 7.2 s longer than ctFS to run Append, and 92% (6.62 s) of that time comes from indexing. The other 8% of the speedup comes from ctFS's improved I/O performance. In contrast, ctFS successfully eliminates most of the overhead of file indexing, primarily by having the MMU perform the indexing in hardware during memory page translation. (See Figure 11 for a breakdown of ctFS's runtime.) It only spends 24 ms in page fault handling, compared to SplitFS's 4.4 s of page fault handling (Section 2). Even though ctFS has a similar number of page faults (525,490) as SplitFS (578,260), SplitFS triggers page faults whose handling requires file indexing, whereas all of ctFS's page faults are minor faults.

For the Append workload, whether running in sync or strict mode does not affect ctFS performance because of ctFS's append optimizations (Section 3.5); ctFS achieves 7.66× speedup over SplitFS in strict mode.

Compared to NOVA's sync mode and pmfs, ctFS-sync achieves 4.4× and 3.87× speedups, respectively.

4.1.2 Other Micro-benchmarks. Figure 10 shows ctFS's performance compared to that of ext4-DAX and SplitFS for the other five microbenchmarks. In sync mode, ctFS achieves an average speedup of 2.17×, 1.97×, 2.43×, 2.97×, against ext4-DAX, SplitFS-sync, pmfs, and NOVA, respectively. In strict mode, ctFS achieves an average speedup of 1.46× and 1.59× against SplitFS-strict and NOVA-strict.

4.1.3 ctFS Runtime Breakdowns. Figure 11 shows the breakdown of ctFS's runtime on each test while running in sync and strict mode, and with huge pages enabled and disabled. We first consider the difference between ctFS's sync and strict modes. Recall that ctFS invokes `pswap` at the end of file overwrite operations under strict mode. This affects both RW and SW. In RW, 68% of the runtime of ctFS-strict is spent on `pswap`. This test represents the worst-case scenario for ctFS-strict, as each write triggers a `pswap` at the smallest granularity (4 KB page): `pswap` cannot perform any optimizations when swapping the entries in the last-level page table, and it is forced to break up the huge pages into base pages.[7] In comparison, while ctFS also needs to invoke `pswap` in SW when running in strict mode, because `pswap` is only invoked once at the end, it incurs negligible overhead (5.7 ms).

[7]Even then, ctFS outperforms SplitFS and NOVA in strict mode, as shown in Figure 10. SplitFS also uses huge pages, so it also needs to break up huge pages, which makes up 37.6% of runtime.

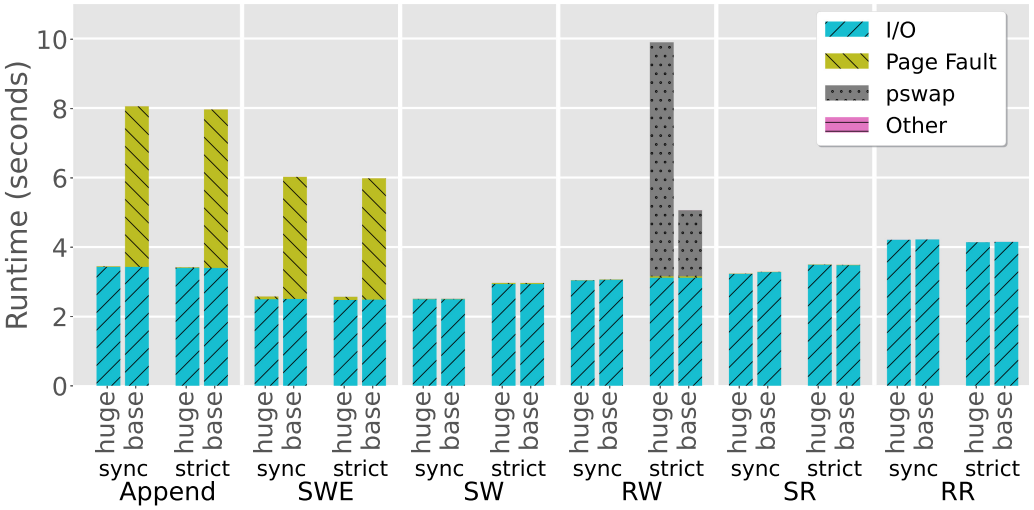

Fig. 11. ctFS overhead breakdown under four configuration combinations: with huge pages enabled and disabled, while running in sync and strict mode.

The figure also shows the difference between having huge pages enabled and disabled. With huge pages enabled, ctFS indeed eliminates much of the indexing overhead, as all workloads are bottlenecked by I/O, except for the RW workload when ctFS runs in strict mode. With huge pages disabled, both the **persistent page table (PPT)** and the DRAM page table have 512× more entries, and each TLB entry now only maps 4 KB instead of 2 MB. For SW, RW, SR, and RR, the overhead after disabling huge pages is negligible in both sync and strict modes (at most 3.4% in SR-strict). This indicates that the overhead of additional TLB misses is negligible. In RR, for example, there are 512× more TLB misses with huge pages disabled, yet this still results in negligible overhead. Note that the number of page faults remains the same even when huge pages are disabled, because ctK copies 512 page table mappings (or the mappings for a 2 MB region) from the PPT to the DRAM page tables on each page fault. In comparison, the large overheads in Append and SWE come from allocating physical PM page frames and building the PPTs, because with only base pages, the PPT contains 512× more entries.

Interestingly, in RW, disabling huge pages results in a 2× speedup for ctFS-strict. This is because with huge pages enabled, every write, which is at the granularity of a base page (4 KB), will trigger a `pswap` that breaks the huge page and causes TLB shootdowns. In comparison, when huge pages are disabled, there is no need to break up huge pages. Therefore, it might be better to allocate base pages at the beginning when there will be a lot of small atomic write operations.

4.2 Real-world Applications

We evaluated ctFS using LevelDB [18] and RocksDB [19], both of which are persistent key-value stores. We drove LevelDB with the **Yahoo! Cloud Serving Benchmark (YCSB)** [4]. YCSB includes six different key-value workloads that are described in Table 2. We drove RocksDB using RocksDB's built-in benchmark `db_bench` with three workloads: *random fill*, which creates and adds key-value pairs; *random read*, which returns the values of given keys; and *random update*, which updates the values of given keys. Each of these tests carries out 5 million operations. Both LevelDB and RocksDB use `pwrite` and `pread` instead of memory-mapped I/O.

Table 2. YCSB Runs and Their Characteristics

A	Update heavy: 50/50 read/write mix
B	Read mostly: 95/5 read/write mix
C	Read only: 100% read
D	Read latest: new records are inserted, and the most recently inserted records are read the most
E	Short ranges: short ranges of records are queried, instead of individual records
F	Read-modify-write: read a record, modify it, and write back the changes

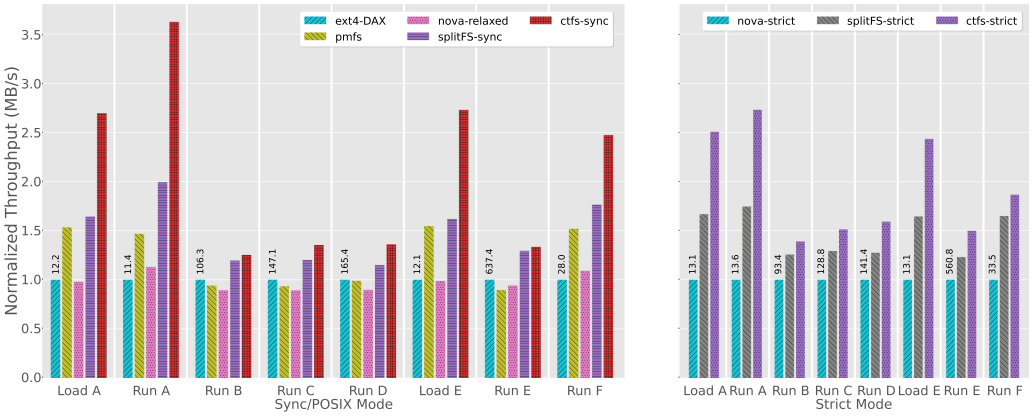

Fig. 12. YCSB on LevelDB. Results are measured in throughput that is normalized to ext4-DAX in the sync/-POSIX group and NOVA-strict in the strict group. The number on top shows the absolute throughput.

The LevelDB workloads demonstrate ctFS's performance advantage achieved by removing the indexing overheads in a real-world application. The RocksDB workloads show that it is feasible and beneficial to replace **write-ahead logs (WALs)** with ctFS's atomic writes.

LevelDB. Figure 12 shows the performance of different PM file systems on LevelDB using the YCSB workloads. ctFS outperforms all the other file systems in each of the workloads when run at comparable consistency levels.

ctFS achieves the most significant improvement in throughput under write-heavy workloads: Load A and E and Run A, B, F.[8] Among these write-heavy workloads, ctFS-sync's throughput is 1.64× the throughput of SplitFS-sync on average, with 1.82× the throughput in the best-case (under Load E). In strict mode, ctFS's throughput is 1.30× higher than that of SplitFS on average, with 1.50× higher in the best-case (under Load A). Compared with ext4-DAX, ctFS-sync has 2.88× higher throughput on average and 3.62× higher throughput in the best case (under Run A).

On read-heavy workloads, ctFS's thoughtput is still higher than that of the other file systems, but by a smaller amount. It achieves an average of 1.25×–1.36× higher throughput over ext4-DAX. As for SplitFS, recall from our microbenchmarks that it spends more time on indexing in random reads than sequential reads. This is why ctFS achieves better throughput than SplitFS in Runs B, C, and D, which are dominated by random reads; e.g., ctFS's throughput is 1.18× and 1.25× higher than that of SplitFS under Run D when run in either sync or strict mode. For Run E, which is dominated by sequential reads, cfFS has 1.02× and 1.22× higher throughput.

By studying the breakdowns of ext4-DAX's time consumption, we observe that indexing time takes up 19.6%, 25%, and 24.5% of the total runtime of LoadA, RunA, and LoadE, respectively.

[8]Load A and Load E create the respective key value stores that are used by the six YCSB runs.

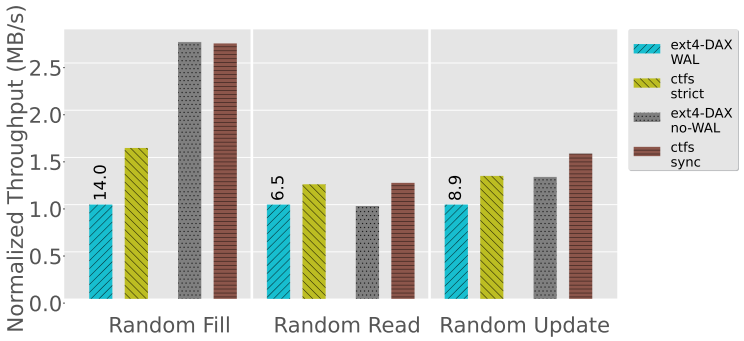

Fig. 13. RocksDB performance.

Table 3. Metadata Operation and Resource Overhead

	SplitFS		Ext4-DAX	ctFS
	sync	strict		
Bootstrap (µs)	1.4×10^6	1.1×10^6	0	5
open (create) (µs)	40	40	15	2
open (existing) (µs)	4	10	4	2
unlink (µs)	32	43	31	1.6
DRAM usage (MB)	198	572	N/A	0.52
Space available after format (GB)	230.7		230.7	248.1
Space used after YCSB LoadA (MB)	5,486		5,337	5,378

There is no difference between sync and strict modes for ctFS.

Meanwhile, ctFS only spends a maximum of 0.2% on indexing overhead (in handling page faults) across all workloads. Hence, indexing accounts for 39.3%, 49.9%, and 46.4% of ctFS's speedup over ext4-DAX on these three workloads. Another 22.5%, 36.4%, and 33% of ctFS's speedups arise from a more efficient I/O data path over ext4-DAX.

RocksDB. We ran our RocksDB experiments two configurations: *strict* and *relaxed*. With *strict*, where data consistency is guaranteed, ext4-DAX is run with WAL enabled, and ctFS is run in strict mode (ctFS-strict) but with WAL disabled. With *relaxed*, where crash consistency is not guaranteed, both ext4-DAX and ctFS-sync are run with WAL disabled.

As shown in Figure 13, With strict, ctFS-strict has 1.60×, 1.22×, and 1.3× the throughput of ext4-DAX for the Random Fill test, the Random Read test and the Random Update test, respectively. This demonstrates the feasibility of replacing WALs in applications with atomic writes in ctFS, as doing so not only improves performance but also simplifies application logic.

With *relaxed*, ctFS-sync is on par with ext4-DAX with the Random Fill test, but has 1.25× and 1.19× the throughput for the Random Read and Random Update test.

4.3 Resource Usage & Other Operations

Table 3 shows the cost of several frequently used file system operations, as well as DRAM overhead after filesystem initialization and space efficiency for ctFS, SplitFS, and Ext4-DAX. Notably, SplitFS spends over one second to initialize because it needs to build the U-Split mapping table, create and `mmap` all the staging files. Similarly, because of the mapping table, SplitFS uses three orders of

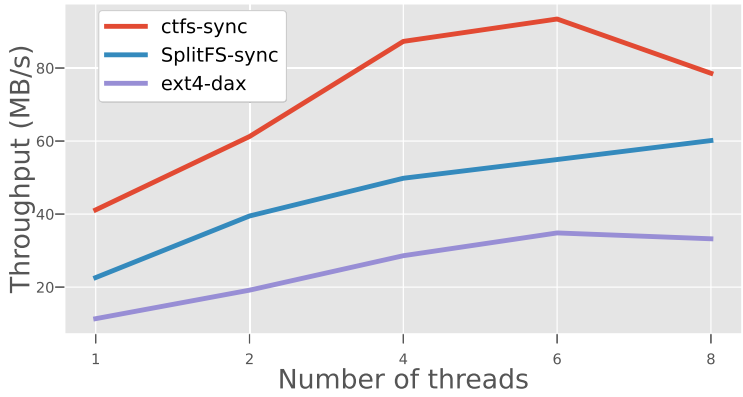

Fig. 14. Scalability of ctFS versus ext4-DAX and SplitFS on LevelDB running YCSB Run A in terms of throughput.

magnitude more DRAM comparing with ctFS. The DRAM usage does not change significantly for SplitFS and ctFS when running workloads, as both of them primarily operate on PM.

In terms of space efficiency, ctFS has 7.52% more available space than ext4-DAX and SplitFS when newly formatted. In fact, ctFS only incurs 10 MB memory overhead for newly formated 248.06 GB space. This is because ctFS allocates inodes and inode bitmaps on demand. After running Load A in the YCSB test on LevelDB, ctFS occupies 0.78% more space than ext4-DAX and 2% less than SplitFS.

Table 3 shows the overhead of file open and unlink of different file systems. Note that both ext4-DAX and SplitFS trap into the kernel, whereas these are user-space operations in ctFS, so the result may not represent a fair comparison. It also shows the bootstrap time: SplitFS spends over 1 second in bootstrap as it memory maps and prefaults the file system, including files, staging files, and unallocated blocks.

Table 3 also shows the memory usage of SplitFS and ctFS. These numbers are measured when the file system is idle, however, they do not change much during the file system execution. SplitFS uses three orders of magnitude more memory, because it needs to create user-space file mapping for staging files.

4.4 Scalability

The design of the cfFS's concurrency model is the same as that of ext4-DAX. Figure 14 shows ctFS's scalability compared with ext4-DAX, running YCSB Run A on LevelDB. All file systems scale similarly. Performance of ctFS peaks at 6 worker threads in an eight-core machine (as 2 additional threads are spawned by LevelDB for other purpose).

5 LIMITATIONS AND DISCUSSION

The design of ctFS presents two unique tradeoffs. First, compared with an in-kernel file system, ctFS's user-space file system design trades protection for performance. While ctFS is not suitable as a general purpose file system, it presents a (rather extreme) tradeoff point for data center applications to squeeze the most out of the hardware, as in data center environments each machine runs only a small number of applications that often trust each other, and protection against intentional attacks is ensured at the boundary of machines or data centers. Furthermore, ctFS can be used as an application's private file system, i.e., where one or several applications own one instance of ctFS. As future work, we plan to implement ctFS entirely in the kernel to present a different

tradeoff point. Additionally, we plan to fork a version in which some of the components are in the kernel for better protection This is made possible because of its extremely low space overhead in formatting (4.3) and the fact that available pages are shared globally. ctFS can also be used in non-sharing mode, so the data of one process will not be mapped to another process's address space. Applications are also recommended to use ctFS in an application-private manner without sharing the data with other (untrusted) applications. However, not protecting against intentional attacks is ctFS's limitation. Nevertheless, if strict protection is needed, then we still provide a number of kernel APIs that can allow permissions to be set for each file.

Second, ctFS's design is also at the expense of internal fragmentation within each fixed-sized partition in the *virtual* memory address space. We argue this is acceptable given the size of today's virtual address spaces. Both Intel and Linux now support 57 bits virtual addresses with 5-level paging, enabling a 128 PB virtual address space. In comparison, the maximum size of Intel Optane DC that can be supported by a server today is 6 TB [29]. Note that ctFS does not waste physical storage space, as any unused regions of a partition are not mapped to physical memory. In addition, ctFS's allocation algorithm is similar to the buddy memory allocator, and hence, the internal fragmentation problem is fundamentally inline with that of modern size-segregated memory allocators such as jemalloc [9], TCMalloc [12], and Go's runtime [15]; the wide adoption of these allocators further suggests that internal fragmentation is an acceptable tradeoff.

6 RELATED WORK

To the best of our knowledge, this article is the first to propose a complete file system that supports contiguous files with a detailed design and evaluation.

SCMFS. SCMFS [39] proposed the high-level idea of allocating each file contiguously in the virtual address space. However, its design is only at a conceptual level. How files are allocated in the virtual memory space is not clearly described. Specifically, it does not address file resizing and external fragmentation, the two fundamental challenges faced by contiguous files. It is unclear what happens if one file expands into the range of another file. Finally, SCMFS's implementation and evaluation are based on using DRAM to simulate PM, instead of actual PM hardware.

File systems for PM. A number of file systems were designed to bypass the kernel. Aerie [37], PMFS [7], Strata [27], SplitFS [23], and ZoFS [6] all allow the user to directly access file data through a user-space component; PMFS, SplitFS, and ZoFS map the metadata and data in application's virtual memory space. In Aerie, metadata updates and locking requests must be sent via IPC to be processed by a trusted system service. Strata logs updates in userspace that are then digested in the kernel. ZoFS strives to provide security by only mapping the metadata to the users who have access permission and only allows trusted library code to modify the metadata by exploiting MPK memory protection keys. Kevin [26], a file system for NAND SSD instead of PM, provides an FPGA implementation of the log-structured merge tree and ports file operations on top.

All of the file systems mentioned above still use a tree-structured index for file indexing. B*e*trFS proposes a B^{ϵ}-tree that is a write-optimized variant B-tree [22]. HashFS [31] uses a global fixed-sized hash table for indexing. However, it still suffers software indexing overhead and its performance is no better when compared to SplitFS. Kuco [2] offloads some indexing from the kernel to the userspace through "collaborative indexing" to improve scalability. However, it still uses traditional ext2-style block mapping. In comparison, ctFS uses a contiguous file design that replaces file indexing.

Crash consistency on file data. Conventional write-ahead logging/journaling [14, 16, 38] typically requires writing the data *twice*: first to journal before updating the target file. The cost of double-write for data may be large, and several mechanisms that avoid data copying have been proposed [1, 3, 17, 28, 30]. Similar to `pswap`, SplitFS's relink is used to efficiently provide atomic writes

without copying the data to the journal. `pswap` differs from `relink` in that the former swaps the virtual-to-physical memory mapping, whereas `relink` changes the mapping within ext4-DAX's extent trees. Failure atomic msync [33] atomically commits changes to a memory mapped file by using ext4's journaling function. SHARE [32] atomically lets pairs of pages share the same physical page in the flash storage. It does not explore the page table hierarchy for optimization.

SubZero [24] proposed a `patch()` function that atomically overwrites the destination region of a mmap file with the content of the source region. `pswap` is different in a few ways. First, `pswap` swaps the mapping, whereas `patch` discards the content in the source region. In addition, `pswap` leverages the page table hierarchy to achieve significant speedup. Finally, `pswap` is mainly used for fast cross-partition expansion and shrink, whereas `patch` is only used for atomic writes.

7 CONCLUDING REMARKS

This article proposes ctFS, a persistent memory file system that offloads file system indexing to the memory management hardware by keeping files contiguous in virtual memory. ctFS is the first such file system built to operate with high efficiency. It carries out most operations in user-space, including the management of the virtual address space, and only maintains virtual-to-physical page mappings in kernel-space. In particular, ctFS gains performance through the use of pswap—an optimized primitive that ensures atomicity while updating page mappings. Our evaluation shows ctFS can outperform ext4-DAX and SplitFS by up to 7.7× and 3.1× and improve YCSB throughputs by up to 3.6× and 1.8×.

In the future, we plan to implement full transaction support, including commit and roll-back, based on `pswap`. Furthermore, the separation design of ctU and ctK makes it possible for applications to create their customized file system layouts on top of ctK. The potential of ctFS is endless.

ACKNOWLEDGMENTS

We thank our shepherd Youngjin Kwon and the anonymous reviewers for their insightful comments. SplitFS authors provided valuable help in obtaining the breakdown of SplitFS's runtime. Rishikesh Devsot and Devin Gibson have ported an early version of ctFS to boot into a Linux shell without hitting Linux's file system.

REFERENCES

[1] Jeff Bonwick and Bill Moore. 2022. ZFS: The Last Word n File Systems. Retrieved from https://wiki.illumos.org/download/attachments/1146951/zfs_last.pdf.

[2] Youmin Chen, Youyou Lu, Bohong Zhu, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, and Jiwu Shu. 2021. Scalable persistent memory file system with Kernel-Userspace collaboration. In *Proceedings of the 19th Conference on File and Storage Technologies*. USENIX Association, 81–95.

[3] Jeremy Condit, Edmund B. Nightingale, Christopher Frost, Engin Ipek, Benjamin Lee, Doug Burger, and Derrick Coetzee. 2009. Better I/O through byte-addressable, persistent memory. In *Proceedings of the 22nd Symposium on Operating Systems Principles (SOSP'09)*. 133–146. DOI: https://doi.org/10.1145/1629575.1629589

[4] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. 2010. Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st Symposium on Cloud Computing (SoCC'10)*. 143–154.

[5] Corbet. 2005. Address space randomization in 2.6. Retrieved from https://lwn.net/Articles/121845/.

[6] Mingkai Dong, Heng Bu, Jifei Yi, Benchao Dong, and Haibo Chen. 2019. Performance and protection in the ZoFS user-space NVM file system. In *Proceedings of the 27th Symposium on Operating Systems Principles (SOSP'19)*. 478–493. DOI: https://doi.org/10.1145/3341301.3359637

[7] Subramanya R. Dulloor, Sanjay Kumar, Anil Keshavamurthy, Philip Lantz, Dheeraj Reddy, Rajesh Sankaran, and Jeff Jackson. 2014. System software for persistent memory. In *Proceedings of the 9th European Conference on Computer Systems (EuroSys'14)*. Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/2592798.2592814

[8] Jake Edge. 2013. Randomizing the kernel. Retrieved from https://lwn.net/Articles/546686/.

[9] Jason Evans. 2006. A scalable concurrent malloc (3) implementation for FreeBSD. In *Proceedings of the of the BSDCAN Conference*.

[10] Alexandra (Sasha) Fedorova. 2019. Why mmap is faster than system calls. Retrieved from https://sasha-f.medium.com/why-mmap-is-faster-than-system-calls-24718e75ab37.
[11] Gregory R. Ganger, Marshall Kirk McKusick, Craig A. N. Soules, and Yale N. Patt. 2000. Soft updates: A solution to the metadata update problem in file systems. *ACM Trans. Comput. Syst.* 18, 2 (May 2000), 127–153. DOI: https://doi.org/10.1145/350853.350863
[12] S. Ghemawat and P. Menage. 2022. TCMalloc. Retrieved from http://goog-perftools.sourceforge.net/doc/.
[13] Mel Gorman. 2022. Page Table Management. Retrieved from https://www.kernel.org/doc/gorman/html/understand/understand006.html.
[14] Jim Gray and Andreas Reuter. 1992. *Transaction Processing: Concepts and Techniques* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA.
[15] Robert Griesemer, Rob Pike, and Ken Thompson. 2022. Golang. Retrieved from https://golang.org/.
[16] R. Hagmann. 1987. Reimplementing the Cedar file system using logging and group commit. In *Proceedings of the 11th ACM Symposium on Operating Systems Principles (SOSP'87)*. 155–162.
[17] Dave Hitz, James Lau, and Michael Malcolm. 1994. File system design for an NFS file server appliance. In *Proceedings of the USENIX Winter 1994 Technical Conference*. USENIX Association.
[18] Google Inc. 2022. LevelDB. Retrieved from https://github.com/google/leveldb.
[19] Meta Platform Inc. 2022. Direct I/O - RocksDB Wiki. Retrieved from https://github.com/facebook/rocksdb/wiki/Direct-IO.
[20] Intel. 2022. Intel AVX-512 Instructions. Retrieved from https://software.intel.com/content/www/us/en/develop/articles/intel-avx-512-instructions.html.
[21] 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. 2019. Basic performance measurements of the Intel Optane DC Persistent Memory. Retrieved from https://arxiv.org/abs/1903.05714v3.
[22] William Jannen, Jun Yuan, Yang Zhan, Amogh Akshintala, John Esmet, Yizheng Jiao, Ankur Mittal, Prashant Pandey, Phaneendra Reddy, Leif Walsh, Michael A. Bender, Martin Farach-Colton, Rob Johnson, Bradley C. Kuszmaul, and Donald E. Porter. 2015. BetrFS: Write-optimization in a kernel file system. *ACM Trans. Storage* 11, 4 (Nov. 2015). DOI: https://doi.org/10.1145/2798729
[23] Rohan Kadekodi, Se Kwon Lee, Sanidhya Kashyap, Taesoo Kim, Aasheesh Kolli, and Vijay Chidambaram. 2019. SplitFS: Reducing software overhead in file systems for persistent memory. In *Proceedings of the 27th Symposium on Operating Systems Principles (SOSP'19)*. 494–508. DOI: https://doi.org/10.1145/3341301.3359631
[24] Juno Kim, Yun Joon Soh, Joseph Izraelevitz, Jishen Zhao, and Steven Swanson. 2020. SubZero: Zero-copy IO for persistent main memory file systems. In *Proceedings of the 11th Asia-Pacific Workshop on Systems (APSys'20)*. Association for Computing Machinery, New York, NY, 1–8. DOI: https://doi.org/10.1145/3409963.3410489
[25] Kenneth C. Knowlton. 1965. A fast storage allocator. *Commun. ACM* 8, 10 (Oct. 1965), 623624. DOI: https://doi.org/10.1145/365628.365655
[26] Jinhyung Koo, Junsu Im, Jooyoung Song, Juhyung Park, Eunji Lee, Bryan S. Kim, and Sungjin Lee. 2021. Modernizing file system through in-storage indexing. In *Proceedings of the 15th Symposium on Operating Systems Design and Implementation (OSDI'21)*. USENIX Association, 75–92. Retrieved from https://www.usenix.org/conference/osdi21/presentation/koo.
[27] Youngjin Kwon, Henrique Fingler, Tyler Hunt, Simon Peter, Emmett Witchel, and Thomas Anderson. 2017. Strata: A cross media file system. In *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP'17)*. 460–477. DOI: https://doi.org/10.1145/3132747.3132770
[28] Eunji Lee, Hyokyung Bahn, and Sam H. Noh. 2013. Unioning of the buffer cache and journaling layers with non-volatile memory. In *Proceedings of the 11th Conference on File and Storage Technologies (FAST'13)*. 73–80.
[29] Lenovo. 2021. Intel Optane Persistent Memory 100 Series Product Guide. Retrieved from https://lenovopress.com/lp1066-intel-optane-persistent-memory-100-series.
[30] C. Mohan. 1999. Repeating history beyond ARIES. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB'99)*. Morgan Kaufmann Publishers Inc., 1–17.
[31] Ian Neal, Gefei Zuo, Eric Shiple, Tanvir Ahmed Khan, Youngjin Kwon, Simon Peter, and Baris Kasikci. 2021. Rethinking file mapping for persistent memory. In *Proceedings of the 19th Conference on File and Storage Technologies (FAST'21)*. USENIX Association, 97–111. Retrieved from https://www.usenix.org/conference/fast21/presentation/neal.
[32] Gihwan Oh, Chiyoung Seo, Ravi Mayuram, YangSuk Kee, and SangWon Lee. 2016. SHARE interface in flash storage for relational and NoSQL databases. In *Proceedings of the International Conference on Management of Data (SIGMOD'16)*. 343–354. DOI: https://doi.org/10.1145/2882903.2882910
[33] Stan Park, Terence Kelly, and Kai Shen. 2013. Failure-atomic Msync(): A simple and efficient mechanism for preserving the integrity of durable data. In *Proceedings of the 8th European Conference on Computer Systems (EuroSys'13)*. 225–238. DOI: https://doi.org/10.1145/2465351.2465374

[34] Soyeon Park, Sangho Lee, Wen Xu, Hyungon Moon, and Taesoo Kim. 2019. Libmpk: Software abstraction for Intel memory protection keys (Intel MPK). In *Proceedings of the USENIX Annual Technical Conference (USENIX-ATC'19)*. 241–254.

[35] Dennis M. Ritchie and Ken Thompson. 1974. The UNIX time-sharing system. *Commun. ACM* 17, 7 (July 1974), 365–375. DOI: https://doi.org/10.1145/361011.361061

[36] Andrew Tanenbaum and Herbert T. Boschung. 2018. *Modern Operating Systems*. Pearson.

[37] Haris Volos, Sanketh Nalli, Sankarlingam Panneerselvam, Venkatanathan Varadarajan, Prashant Saxena, and Michael M. Swift. 2014. Aerie: Flexible file-system interfaces to storage-class memory. In *Proceedings of the 9th European Conference on Computer Systems (EuroSys'14)*. DOI: https://doi.org/10.1145/2592798.2592810

[38] David Woodhouse. 2001. JFFS : The journalling flash file system. In *Proceedings of the Ottawa Linux Symposium*. RedHat Inc.

[39] Xiaojian Wu and A. L. Narasimha Reddy. 2011. SCMFS: A file system for storage class memory. In *Proceedings of the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC'11)*. 1–11.

[40] Jian Xu and Steven Swanson. 2016. NOVA: A log-structured file system for hybrid volatile/non-volatile main memories. In *Proceedings of the 14th Conference on File and Storage Technologies (FAST'16)*. 323–338. Retrieved from https://www.usenix.org/conference/fast16/technical-sessions/presentation/xu.

Received 20 June 2022; revised 20 June 2022; accepted 13 September 2022