Rethink the Sync

EDMUND B. NIGHTINGALE, KAUSHIK VEERARAGHAVAN,
PETER M. CHEN, and JASON FLINN
University of Michigan

We introduce *external synchrony*, a new model for local file I/O that provides the reliability and simplicity of synchronous I/O, yet also closely approximates the performance of asynchronous I/O. An external observer cannot distinguish the output of a computer with an externally synchronous file system from the output of a computer with a synchronous file system. No application modification is required to use an externally synchronous file system. In fact, application developers can program to the simpler synchronous I/O abstraction and still receive excellent performance. We have implemented an externally synchronous file system for Linux, called xsyncfs. Xsyncfs provides the same durability and ordering-guarantees as those provided by a *synchronously* mounted ext3 file system. Yet even for I/O-intensive benchmarks, xsyncfs performance is within 7% of ext3 mounted *asynchronously*. Compared to ext3 mounted synchronously, xsyncfs is up to two orders of magnitude faster.

Categories and Subject Descriptors: D.4.3 [**Operating Systems**]: File Systems Management; D.4.7 [**Operating Systems**]: Organization and Design; D.4.8 [**Operating Systems**]: Performance

General Terms: Performance, Design

Additional Key Words and Phrases: File systems, synchronous I/O, speculative execution, causality

ACM Reference Format:

Nightingale, E. B., Kaushik, V., Chen, P. M., and Flinn, J. 2008. Rethink the Sync. ACM Trans. Comput. Syst. 26, 3, Article 6 (September 2008), 26 pages. DOI=10.1145/1394441.1394442 http://doi.acm.org/10.1145/1394441.1394442

This work has been supported by the National Science Foundation under award CNS-0509093. Jason Flinn is supported by NSF CAREER award CNS-0346686, and Ed Nightingale is supported by a Microsoft Research Student Fellowship. Intel Corp. has provided additional support. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NSF, Intel, Microsoft, or the University of Michigan.

Authors' addresses: E. B. Nightingale, Microsoft Research, Redmond, WA 98052; email: ed.nightingale@microsoft.com; K. Veeraraghavan, P. M. Chen, and J. Flinn, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122; K. Veeraraghavan; email: kaushikv@eecs.umich.edu; P. M. Chen and J. Flinn; email: {pmchen, jflinn}@umich.edu.

1. INTRODUCTION

File systems serve two opposing masters: durability and performance. The tension between these goals has led to two models of file I/O: synchronous and asynchronous.

A synchronous file system (e.g., one mounted with the `sync` option on a Linux system) guarantees durability by blocking the calling application until modifications are committed to disk. Synchronous I/O provides a clean abstraction to users. Any file system operation that is visible to the user as having completed is durable—data will not be lost due to a subsequent OS crash or power failure. Synchronous I/O also guarantees the ordering of modifications; if one operation causally precedes another, the effects of the second operation are never visible unless the effects of first operation are also visible. Unfortunately, synchronous I/O can be very slow because applications are frequently blocked while waiting for mechanical disk operations. In fact, our results show that blocking due to synchronous I/O can degrade the performance of disk-intensive benchmarks by two orders of magnitude.

In contrast, an asynchronous file system does not block the calling application, so modifications are typically committed to disk long after the call completes. This is fast, but not safe. Users view output that depends on uncommitted modifications. If the system crashes or loses power before those modifications commit, the output observed by the user will be invalid. Asynchronous I/O also complicates applications that require durability or ordering-guarantees. Programmers must insert explicit synchronization operations, such as `fsync`, to enforce the guarantees required by their applications. They must sometimes implement complex group-commit strategies to achieve reasonable performance. Despite the poor guarantees provided to users and programmers, most local file systems provide an asynchronous I/O abstraction by default, because synchronous I/O is simply too slow.

The tension between durability and performance leads to surprising behavior. For instance, on most desktop operating systems, even executing an explicit synchronization command, such as `fsync`, does not protect against data loss in the event of a power failure [McKusick 2006]. This behavior is not a bug, but rather a conscious design decision to sacrifice durability for performance [Slashdot 2005]. For example, on `fsync`, the Linux 2.4 kernel commits data to the volatile hard drive cache rather than to the disk platter. If a power failure occurs, the data in the drive cache is lost. Because of this behavior, applications that require stronger durability guarantees, such as the MySQL database, recommend disabling the drive cache [MySQL AB 2006]. While MacOS X and the Linux 2.6 kernel provide mechanisms to explicitly flush the drive cache, these mechanisms are not enabled by default, due to the severe performance degradation they can cause.

We show that a new model of local file I/O, which we term *external synchrony*, resolves the tension between durability and performance. External synchrony provides the reliability and simplicity of synchronous I/O, while closely approaching the performance of asynchronous I/O. In external synchrony, we view the abstraction of synchronous I/O as a set of guarantees that are provided to

the clients of the file system. In contrast to asynchronous I/O, which improves performance by substantially weakening these guarantees, externally synchronous I/O provides the same guarantees, but it changes the clients to which the guarantees are provided.

Synchronous I/O reflects the *application-centric* view of modern operating systems. The return of a synchronous file system call guarantees durability to the application, since the calling process is blocked until modifications commit. In contrast, externally synchronous I/O takes a *user-centric* view in which it guarantees durability not to the application, but to any external entity that observes application output. An externally synchronous system returns control to the application before committing data. However, it subsequently buffers all output that causally depends on the uncommitted modification. Buffered output is only externalized (sent to the screen, network, or other external device) after the modification commits.

From the viewpoint of an external observer such as a user or an application running on another computer, the guarantees provided by externally synchronous I/O are identical to the guarantees provided by a traditional file system mounted synchronously. An external observer never sees output that depends on uncommitted modifications. Since external synchrony commits modifications to disk in the order in which they are generated by applications, an external observer will not see a modification unless all other modifications that causally precede that modification are also visible. However, because externally synchronous I/O rarely blocks applications, its performance approaches that of asynchronous I/O.

Our externally synchronous Linux file system, xsyncfs, uses mechanisms developed as part of the Speculator project [Nightingale et al. 2006]. When a process performs a synchronous I/O operation, xsyncfs validates the operation, adds the modifications to an ext3 file system transaction, and returns control to the calling process without waiting for the transaction to commit. However, xsyncfs also taints the calling process with a *commit dependency* that specifies that the process is not allowed to externalize any output until the transaction commits. If the process writes to the network, screen, or other external device, its output is buffered by the operating system. The buffered output is released only after all disk transactions on which the output depends commit. If a process with commit dependencies interacts with another process on the same computer through IPC, such as pipes, the file cache, or shared memory, the other process inherits those dependencies so that it also cannot externalize output until the transaction commits. The performance of xsyncfs is generally quite good, since applications can perform computation and initiate further I/O operations while waiting for a transaction to commit. In most cases, output is delayed by no more than the time to commit a single transaction—this is typically less than the perception threshold of a human user.

Xsyncfs uses *output-triggered commits* to balance throughput and latency. Output-triggered commits track the causal relationship between external output and file system modifications, to decide when to commit data. Until some external output is produced that depends upon modified data, xsyncfs may delay committing data to optimize for throughput. However, once some output

is buffered that depends upon an uncommitted modification, an immediate commit of that modification is triggered to minimize latency for any external observer.

Our results to date are very positive. For I/O-intensive benchmarks such as Postmark and an Andrew-style build, the performance of xsyncfs is within 7% of the default asynchronous implementation of ext3. Compared to current implementations of synchronous I/O in the Linux kernel, external synchrony offers better performance and better reliability. Xsyncfs is as much as an order of magnitude faster than the default version of ext3 mounted synchronously, which allows data to be lost on power failure because committed data may reside in the volatile hard drive cache. Xsyncfs is as much as two orders of magnitude faster than a version of ext3 that guards against losing data on power failure. Xsyncfs sometimes even improves the performance of applications that do their own custom synchronization. Running on top of xsyncfs, the MySQL database executes a modified version of the TPC-C benchmark up to three times faster than when it runs on top of ext3 mounted asynchronously.

2. DESIGN OVERVIEW

2.1 Principles

The design of external synchrony is based on two principles. First, we define externally synchronous I/O by its externally observable behavior rather than by its implementation. Second, we note that application state is an internal property of the computer system. Since application state is not directly observable by external entities, the operating system need not treat changes to application state as an external output.

Synchronous I/O is usually defined by its implementation: an I/O is considered synchronous if the calling application is blocked until after the I/O completes [Silberschatz and Galvin 1998]. In contrast, we define externally synchronous I/O by its observable behavior: we say that an I/O is externally synchronous if the external output produced by the computer system cannot be distinguished from output that could have been produced if the I/O had been synchronous.

The next step is to precisely define what is considered external output. Traditionally, the operating system takes an *application-centric* view of the computer system, in which it considers applications to be external entities observing its behavior. This view divides the computer system into two partitions: the kernel, which is considered internal state, and the user level, which is considered external state. Using this view, the return from a system call is considered an externally visible event.

Users, not applications, however, are the true observers of the computer system. Application state is only visible through output sent to external devices such as the screen and network. By regarding application state as internal to the computer system, the operating system can take a *user-centric* view in which only output sent to an external device is considered externally visible. This view divides the computer system into three partitions, the kernel and applications,

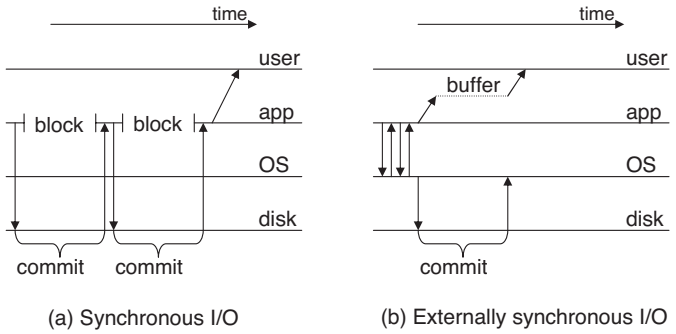


Fig. 1. Example of externally synchronous file I/O. This figure shows the behavior of a sample application that makes two file system modifications, then displays output to an external device. The diagram on the left shows how the application executes when its file I/O is synchronous; the diagram on the right shows how it executes when its file I/O is externally synchronous.

both of which are considered internal state, and the external interfaces, which are considered externally visible. Using this view, changes to application state, such as the return from a system call, are not considered externally visible events.

The operating system can implement user-centric guarantees because it controls access to external devices. Applications can only generate external events with the cooperation of the operating system. Applications must invoke this cooperation either directly, by making a system call, or indirectly, by mapping an externally visible device.

2.2 Correctness

Figure 1 illustrates these principles by showing an example single-threaded application that makes two file system modifications and writes some output to the screen. In Figure 1(a), the file modifications made by the application are synchronous. Thus the application blocks until each modification commits.

We say that external output of an externally synchronous system is equivalent to the output of a synchronous one if (a) the values of the external outputs are the same, and (b) the outputs occur in the same causal order, as defined by Lamport's *happens before* relation [1978]. We consider disk-commits to be external output because they change the stable image of the file system. If the system crashes and reboots, the change to the stable image is visible. Since the operating system cannot control when crashes occur, it must treat disk commits as external output. Thus, in Figure 1(a), there are three external outputs: the two commits and the message displayed on the screen.

An externally synchronous file I/O returns the same result to applications that would have been returned by a synchronous I/O. The file system does all processing that would be done for a synchronous I/O, including validation and changing the volatile (in-memory) state of the file system, except that it does not actually commit the modification to disk before returning. Because the results that an application sees from an externally synchronous I/O are equivalent to

the results it would have seen if the I/O had been synchronous, the external output it produces is the same in both cases.

An operating system that supports external synchrony must ensure that external output occurs in the same causal order in which it would have occurred had I/O been performed synchronously. Specifically, if an external output causally follows an externally synchronous file I/O, then that output cannot be observed before the file I/O has been committed to disk. In the example, this means that the second file modification made by the application cannot commit before the first, and that the screen output cannot be seen before both modifications commit.

2.3 Improving Performance

The externally synchronous system in Figure 1(b) makes two optimizations to improve performance. First, the two modifications are group-committed as a single file system transaction. Because the commit is atomic, the effects of the second modification are never seen unless the effects of the first are also visible. Grouping multiple modifications into one transaction has many benefits: the commit of all modifications is done with a single sequential disk write, writes to the same disk block are coalesced in the log, and no blocks are written to disk at all if data writes are closely followed by deletion. For example, ext3 employs value logging—when a transaction commits, only the latest version of each block is written to the journal. If a temporary file is created and deleted within a single transaction, none of its blocks are written to disk. In contrast, a synchronous file system cannot group multiple modifications for a single-threaded application because the application does not begin the second modification until after the first commits.

The second optimization is buffering screen output. The operating system must delay the externalization of the output until after the commit of the file modifications in order to obey the causal ordering constraint of externally synchronous I/O. One way to enforce this ordering would be to block the application when it initiates external output. However, the asynchronous nature of the output enables a better solution. The operating system instead buffers the output and allows the process that generated the output to continue execution. After the modifications are committed to disk, the operating system releases the output to the device for which it was destined.

This design requires that the operating system track the causal relationship between file system modifications and external output. When a process writes to the file system, it inherits a commit dependency on the uncommitted data that it wrote. When a process with commit dependencies modifies another kernel object (process, pipe, file, UNIX socket, etc.) by executing a system call, the operating system marks the modified objects with the same commit dependencies. Similarly, if a process observes the state of another kernel object with commit dependencies, the process inherits those dependencies. If a process with commit dependencies executes a system call for which the operating system cannot track the flow of causality (e.g., an `ioctl`), the process is blocked until its file system modifications have been committed. Any external output

inherits the commit dependencies of the process that generated it—the operating system buffers the output until the last dependency is resolved by committing modifications to disk.

2.4 Deciding When to Commit

An externally synchronous file system uses the causal relationship between external output and file modifications to trigger commits. There is a well-known tradeoff between throughput and latency for group commit strategies. Delaying a group commit in the hope that more modifications will occur in the near future can improve throughput by amortizing more modifications across a single commit. However, delaying a commit also increases latency—in our system, commit latency is especially important because output cannot be externalized until the commit occurs.

Latency is unimportant if no external entity is observing the result. Specifically, until some output is generated that causally depends on a file system transaction, committing the transaction does not change the observable behavior of the system. Thus, the operating system can improve throughput by delaying a commit until some output that depends on the transaction is buffered (or until some application that depends on the transaction blocks due to an `ioctl` or similar system call). We call this strategy *output-triggered commits* since the attempt to generate output that is causally dependent upon modifications to be written to disk triggers the commit of those modifications.

Output-triggered commits enable an externally synchronous file system to maximize throughput when output is not being displayed (for example, when it is piped to a file). However, when a user could be actively observing the results of a transaction, commit latency is small.

2.5 Limitations

One potential limitation of external synchrony is that it complicates application-specific recovery from catastrophic media failure, because the application continues execution before such errors are detected. Although the kernel validates each modification before writing it to the file cache, the physical write of the data to disk may subsequently fail. While smaller errors such as a bad disk block are currently handled by the disk or device driver, a catastrophic media failure is rarely masked at these levels. Theoretically, a file system mounted synchronously could propagate such failures to the application. However, a recent survey of common file systems [Prabhakaran et al. 2005] found that write errors are either not detected by the file system (ext3, jbd, and NTFS) or induce a kernel panic (ReiserFS). An externally synchronous file system could propagate failures to applications by using Speculator to checkpoint a process before it modifies the file system. If a catastrophic failure occurs, the process would be rolled back and notified of the failure. We rejected this solution because it would both greatly increase the complexity of external synchrony and severely penalize its performance. Further, it is unclear that catastrophic failures are best handled by applications—it seems best to handle them in the

operating system, either by inducing a kernel panic or (preferably) by writing data elsewhere.

Another limitation of external synchrony is that the user may have some temporal expectations about when modifications are committed to disk. As defined so far, an externally synchronous file system could indefinitely delay committing data written by an application with no external output. If the system crashes, a substantial amount of work could be lost. Xsyncfs therefore commits data every five seconds, even if no output is produced. The five second commit interval is the same value used by ext3 mounted asynchronously.

A final limitation of external synchrony is that modifications to data in two different file systems cannot be easily committed with a single disk transaction. Potentially, we could share a common journal among all local file systems, or we could implement a two-phase commit strategy. However, a simpler solution is to block a process with commit dependencies for one file system before it modifies data in a second. Speculator would map each dependency to a specific file system. When a process writes to a file system, Speculator would verify that the process depends only on the file system it is modifying; if it depends on another file system, Speculator would block it until its previous modifications commit.

3. IMPLEMENTATION

External synchrony requires that an operating system implement two mechanisms: tracking the causal dependencies of applications, and buffering visible output. These mechanisms, when used by an externally synchronous file system, allow the release of visible output dependent on pending disk write to be deferred until the writes complete. Each mechanism is already available as part of Speculator [Nightingale et al. 2006], which provides operating system support for multi-process speculative execution. Therefore, we begin by describing our implementation of operating system support for external synchrony within the context of Speculator.

3.1 OS Support for External Synchrony

Speculator adds two new data structures to the kernel. A *speculation* object tracks all process and kernel state that depend on the success or failure of a speculative operation. Each speculative object in the kernel has an *undo log* that contains the state needed to undo speculative modifications to that object. As processes interact with kernel objects by executing system calls, Speculator uses these data structures to track causal dependencies. For example, if a speculative process writes to a pipe, Speculator creates an entry in the pipe's undo log that refers to the speculations on which the writing process depends. If another process reads from the pipe, Speculator creates an undo log entry for the reading process that refers to all speculations on which the pipe depends.

In external synchrony, a commit dependency represents the causal relationship between kernel state and an uncommitted file system modification. Any kernel object that has one or more associated commit dependencies is referred to as *uncommitted*. Any external output from a process that is uncommitted is

buffered within the kernel until the modifications on which the output depends have been committed. This ensures that uncommitted output is never visible to an external observer.

When a process writes to an externally synchronous file system, Speculator marks the process as uncommitted. It also creates a commit dependency between the process and the uncommitted file system transaction that contains the modification. When the file system commits the transaction to disk, the commit dependency is removed. Once all commit dependencies for buffered output have been removed, Speculator releases that output to the external device to which it was written. When the last commit dependency for a process is discarded, Speculator marks the process as committed.

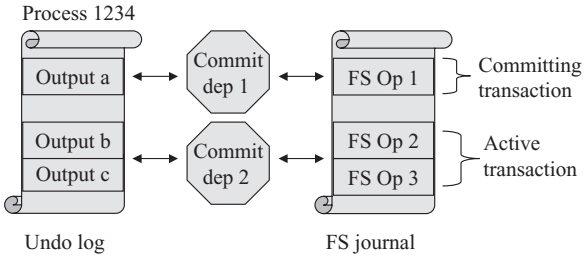
Speculator propagates commit dependencies among kernel objects and processes using the same mechanisms it uses to propagate speculative dependencies. Speculator also maintains the same many-to-many relationship between commit dependencies and undo logs as it does for speculations and undo logs. Since commit dependencies are never rolled back, undo logs need not contain data to undo the effects of an operation. Therefore, undo logs in an externally synchronous system only track the relationship between commit dependencies and kernel objects, and reveal which buffered output can be safely released. This simplicity enables Speculator to support more forms of interaction among uncommitted processes than it supports for speculative processes. For example, checkpointing multi-threaded processes for speculative execution is a thorny problem [Nightingale et al. 2006; Qin et al. 2005]. However, as discussed in Section 3.6, tracking their commit dependencies is substantially simpler.

3.2 File System Support for External Synchrony

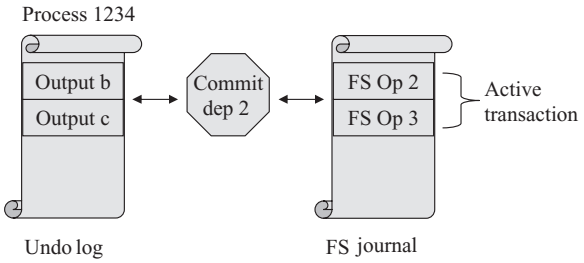
An externally synchronous file system must commit modifications to disk in causal order. Causally ordered writes ensure that modifications are written to disk in the same order in which they would have been written using synchronous I/O.

Which underlying data structure is appropriate to provide ordering equivalent to synchronous I/O? A file system must provide some way to order writes as they are sent to disk, and it must have a notification mechanism to determine when a write completes such that it is safe to release buffered output. These two properties are provided by either a logging file system or a file system journal. Note that a logging file system, such as LFS, or a journaling file system, such as ext3, provides stronger guarantees than are required of synchronous I/O. Synchronous I/O requires only that modifications are causally ordered; it requires neither atomicity of writes nor a mechanism to prevent disk corruption or data loss when a crash occurs. If a file system such as ext2 were to order writes in memory and then update files in-place, it could be converted to an externally synchronous file system. We chose to use journaling to provide ordering because it is widely available and already used by ext3, which we modified to create xsyncfs.

Ordering must also be provided at the device interface, lest the disk reorder writes and break the invariants defined by the file system. Device layer ordering



(a) Data structures with a committing and active transaction



(b) Data structures after the first transaction commits

Fig. 2. The external synchrony data structures.

is provided on some disks through tagged command queuing (TCQ). For disks that do not support TCQ, ordering can be achieved by disabling the drive cache (and executing writes one at a time), putting the drive cache in write-through mode, or implementing an abstraction, such as write barriers, that flushes the drive cache to disk.

3.3 Xsyncfs

We modified ext3, a journaling Linux file system, to create xsyncfs. In its default *ordered* mode, ext3 writes only metadata modifications to its journal. In its *journaled* mode, ext3 writes both data and metadata modifications. Modifications from many different file system operations may be grouped into a single compound journal transaction that is committed atomically. Ext3 writes modifications to the *active* transaction—at most one transaction may be active at any given time. A commit of the active transaction is triggered when journal space is exhausted, an application performs an explicit synchronization operation such as `fsync`, or the oldest modification in the transaction is more than five seconds old. After the transaction starts to commit, the next modification triggers the creation of a new active transaction. Only one transaction may be committing at any given time, so the next transaction must wait for the commit of the prior transaction to finish before it commits.

Figure 2 shows how the external synchrony data structures change when a process interacts with xsyncfs. In Figure 2(a), process 1234 has completed three file system operations, sending output to the screen after each one. Since the output after the first operation has triggered a transaction commit, the two following operations have been placed in a new active transaction. The output

is buffered in the undo log; the commit dependencies maintain the relationship between buffered output and uncommitted data. In Figure 2(b), the first transaction has been committed to disk. Therefore, the output that depended upon the committed transaction has been released to the screen and the commit dependency has been discarded.

Xsyncfs uses journaled mode rather than the default ordered mode. This change guarantees ordering—the property requiring that if an operation A causally precedes another operation, B, the effects of B should never be visible unless the effects of A are also visible. This guarantee requires that B never be committed to disk before A. Otherwise, a system crash or power failure could occur between the two commits—in this case, after the system is restarted, B would be visible when A is not. Since journaled mode adds all modifications for A to the journal before the operation completes, those modifications must already be in the journal when B begins (since B causally follows A). Thus, either B is part of the same transaction as A (in which case the ordering property holds, since A and B are committed atomically), or the transaction containing A is already committed before the transaction containing B starts to commit.

In contrast, the default mode in ext3 does not provide ordering, since data modifications are not journaled. The kernel may write the dirty blocks of A and B to disk in any order as long as the data reaches disk before the metadata in the associated journal transaction commits. Thus the data modifications for B may be visible after a crash without the modifications for A being visible.

Xsyncfs informs Speculator when a new journal transaction is created—this allows Speculator to track state that depends on the uncommitted transaction. Xsyncfs also informs Speculator when a new modification is added to the transaction and when the transaction commits.

As described in Section 1, the default behavior of ext3 does not guarantee that modifications are durable after a power failure. In the Linux 2.4 kernel, durability can be ensured only by disabling the drive cache. The Linux 2.6.11 kernel provides the option of using write barriers to flush the drive cache before and after writing each transaction commit record. Since Speculator runs on a 2.4 kernel, we ported write barriers to our kernel and modified xsyncfs to use write barriers to guarantee that all committed modifications are preserved, even on power failure.

3.4 Output-Triggered Commits

Xsyncfs uses the causal relationship between disk I/O and external output to balance the competing concerns of throughput and latency. Currently, ext3 commits its journal every five seconds, which typically groups the commits of many file system operations. This strategy optimizes for throughput, a logical behavior when writes are asynchronous. However, latency is an important consideration in xsyncfs, since users must wait to view output until the transactions on which that output depends commit. If xsyncfs, were to use the default ext3 commit strategy, disk throughput would be high, but the user might be forced to wait up to five seconds to see output. This behavior is clearly unacceptable for interactive applications.

We therefore modified Speculator to support output-triggered commits. Speculator provides callbacks to xsyncfs when it buffers output or blocks a process that performed a system call for which it cannot track the propagation of causal dependencies (e.g., an `ioctl`). Xsyncfs uses the ext3 strategy of committing every five seconds unless it receives a callback that indicates that Speculator blocked or buffered output from a process that depends on the active transaction. The receipt of a callback triggers a commit of the active transaction.

Output-triggered commits adapt the behavior of the file system according to the observable behavior of the system. For instance, if a user directs output from a running application to the screen, latency is reduced by committing transactions frequently. If the user instead redirects the output to a file, xsyncfs optimizes for throughput by committing every five seconds. Optimizing for throughput is correct in this instance, since the only event the user can observe is the completion of the application (and the completion would trigger a commit if it is a visible event). Finally, if the user were to observe the contents of the file using a different application, for example, `tail`, xsyncfs would correctly optimize for latency because Speculator would track the causal relationship through the kernel data structures from `tail` to the transaction, and provide callbacks to xsyncfs. When `tail` attempts to output data to the screen, Speculator callbacks will cause xsyncfs to commit the active transaction.

3.5 Rethinking Sync

Asynchronous file systems provide explicit synchronization operations, such as `sync` and `fdatasync`, for applications with durability or ordering constraints. In a synchronous file system, such synchronization operations are redundant, since ordering and durability are already guaranteed for all file system operations. However, in an externally synchronous file system, some extra support is needed to minimize latency. For instance, a user who types "sync" in a terminal would prefer that the command complete as soon as possible.

When xsyncfs receives a synchronization call such as `sync` from the VFS layer, it creates a commit dependency between the calling process and the active transaction. Since this does not require a disk write, the return from the synchronization call is almost instantaneous. If a visible event occurs, such as the completion of the `sync` process, Speculator will issue a callback that causes xsyncfs to commit the active transaction.

External synchrony simplifies the file system abstraction. Since xsyncfs requires no application modification, programmers can write the same code that they would write if they were using an unmodified file system mounted synchronously. They do not need explicit synchronization calls to provide ordering and durability, since xsyncfs provides these guarantees by default for all file system operations. Further, since xsyncfs does not incur the large performance penalty usually associated with synchronous I/O, programmers do not need complicated group-commit strategies to achieve acceptable performance. Group-commit is provided transparently by xsyncfs.

Of course, a hand-tuned strategy might offer better performance than the default policies provided by xsyncfs. However, as described in Section 3.4, there are

some instances in which xsyncfs can optimize performance when an application solution cannot. Since xsyncfs uses output-triggered commits, it knows when no external output has been generated that depends on the current transaction; in these instances, xsyncfs uses group-commit to optimize throughput. In contrast, an application-specific commit strategy cannot determine the visibility of its actions beyond the scope of the currently executing process; it must therefore conservatively commit modifications before producing external messages.

For example, consider a client that issues two sequential transactions to a database server on the same computer, and then produces output. Xsyncfs can safely group the commit of both transactions. However, the database server (which does not use output-triggered commits) must commit each transaction separately, since it cannot know whether or not the client will produce output after it is informed of the commit of the first transaction.

3.6 Shared Memory

Speculator does not propagate speculative dependencies when processes interact through shared memory, due to the complexity of checkpointing at arbitrary states in the execution of a process. Since commit dependencies do not require checkpoints, we enhanced Speculator to propagate them among processes that share memory.

Speculator can track causal dependencies because processes can only interact through the operating system. Usually, this interaction involves an explicit system call (e.g., `write`) that Speculator can intercept. However, when processes interact through shared memory regions, only the sharing and unsharing of regions is visible to the operating system. Thus Speculator cannot readily intercept individual reads and writes to shared memory.

We considered marking a shared memory page inaccessible when a process with write permission inherits a commit dependency that a process with read permission does not have. This would trigger a page fault whenever a process reads or writes the shared page. If a process reads the page after another writes it, any commit dependencies would be transferred from the writer to the reader. Once these processes have the same commit dependencies, the page can be restored to its normal protections. We felt this mechanism would perform poorly because of the time needed to protect and unprotect pages, as well as the extra page faults that would be incurred.

Instead, we decided to use an approach that imposes less overhead but might transfer dependencies when not strictly necessary. We make a conservative assumption that processes with write permission for a shared memory region are continually writing to that region, while processes with read permission are continually reading it. When a process with write permission for a shared region inherits a new commit dependency, any process with read permission for that region atomically inherits the same dependency.

Speculator uses the same mechanism to track commit dependencies transfered through memory-mapped files. Similarly, Speculator is conservative when propagating dependencies for multi-threaded applications—any dependency inherited by one thread is inherited by all.

4. EVALUATION

Our evaluation answers the following questions:

—How does the durability of xsyncfs compare to current file systems?
—How does the performance of xsyncfs compare to current file systems?
—How does xsyncfs affect the performance of applications that synchronize explicitly?
—How much do output-triggered commits improve the performance of xsyncfs?

4.1 Methodology

All computers used in our evaluation have a 3.02 GHZ Pentium 4 processor with 1 GB of RAM. Each computer has a single Western Digital WV-XL40 hard drive, which is a 7200 RPM 120 GB ATP 100 drive with a 2 MB on-disk cache. The computers run Red Hat Enterprise Linux version 3 (kernel version 2.4.21). We use a 400 MB journal size for both ext3 and xsyncfs. For each benchmark, we measured ext3 executing in both journaled and ordered mode. Since journaled mode executed faster in every benchmark, we report only journaled mode results in this evaluation. We note that ordered mode is faster than journaled mode when a very large sequential write fills the journal. Performance suffers because the process writing into the file system is blocked while the journal is truncated. None of our benchmarks contain sequential writes large enough to trigger this behavior. Finally, we measured the performance of ext3 both using write barriers and with the drive cache disabled. In all cases, write barriers were faster than disabling the drive cache, since the drive cache improves read times and reduces the frequency of writes to the disk platter. Thus, we report only results using write barriers.

4.2 Durability

Our first benchmark empirically confirms that without write barriers, ext3 does not guarantee durability. This result holds in both journaled and ordered mode, whether ext3 is mounted synchronously or asynchronously, and even if `fsync` commands are issued by the application after every write. Even worse, our results show that, despite the use of journaling in ext3, a loss of power can corrupt data and metadata stored in the file system.

We confirmed these results by running an experiment in which a test computer continuously writes data to its local file system. After each write completes, the test computer sends a UDP message that is logged by a remote computer. During the experiment, we cut power to the test computer. After it reboots, we compare the state of its file system to the log on the remote computer.

Our goal was to determine when each file system guarantees durability and ordering. We say a file system fails to provide durability if the remote computer logs a message for a write operation, but the test computer is missing the data written by that operation. In this case, durability is not provided because an external observer (the remote computer) saw output that depended on data that was subsequently lost. We say a file system fails to provide ordering if the state of the file after reboot violates the temporal ordering of writes. Specifically, for

File system configuration	Data durable on `write`	Data durable on `fsync`
Asynchronous	No	Not on power failure
Synchronous	Not on power failure	Not on power failure
Synchronous with write barriers	Yes	Yes
External synchrony	Yes	Yes

Fig. 3. When is data safe? This figure describes whether each file system provides durability to the user when an application executes a `write` or `fsync` system call. A "Yes" indicates that the file system provides durability if an OS crash or power failure occurs.

each block in the file, ordering is violated if the file does not also contain all previously-written blocks.

For each configuration shown in Figure 3, we ran four trials of this experiment: two in journaled mode, and two in ordered mode. As expected, our results confirm that ext3 provides durability only when write barriers are used. Without write barriers, synchronous operations ensure only that modifications are written to the hard drive cache. If power fails before the modifications are written to the disk platter, those modifications are lost.

Some of our experiments exposed a dangerous behavior in ext3: unless write barriers are used, power failures can corrupt both data and metadata stored on disk. In one experiment, a block in the file being modified was silently overwritten with garbage data. In another, a substantial amount of metadata in the file system, including the superblock, was overwritten with garbage. In the latter case, the test machine failed to reboot until the file system was manually repaired. In both cases, corruption was caused by the commit block for a transaction being written to the disk platter before all data blocks in that transaction were written to disk. Although the operating system wrote the blocks to the drive cache in the correct order, the hard drive reorders the blocks when writing them to the disk platter. After this happens, the transaction is committed during recovery, even though several data blocks do not contain valid data, effectively overwriting disk blocks with uninitialized data.

Our results also confirm that ext3 without write barriers writes data to disk out of order. Journaled mode alone is insufficient to provide ordering, since the order of writing transactions to the disk platter may differ from the order of writing transactions to the drive cache. In contrast, ext3 provides both durability and ordering when write barriers are combined with some form of synchronous operation (either mounting the file system synchronously or calling `fsync` after each modification). If write barriers are not available, the equivalent behavior could also be achieved by disabling the hard drive cache.

The last row of Figure 3 shows results for xsyncfs. As expected, xsyncfs provides both durability and ordering.

4.3 The PostMark Benchmark

We next ran the PostMark benchmark, which was designed to replicate the small file workloads seen in electronic mail, netnews, and Web based commerce [Katcher 1997]. We used PostMark version 1.5, running in a configuration that creates 10,000 files, performs 10,000 transactions consisting of file

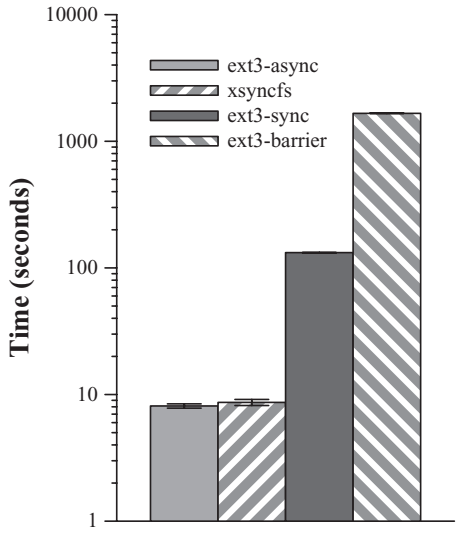


Fig. 4. The PostMark file system benchmark. This figure shows the time to run the PostMark benchmark—the y-axis is logarithmic. Each value is the mean of five trials—the relatively small error bars are 90% confidence intervals.

reads, writes, creates, and deletes, and then removes all files. File sizes ranged from 500 to 10,000 bytes. The PostMark benchmark has a single thread of control that executes file system operations as quickly as possible. PostMark is a good test of file system throughput, since it does not generate any output or perform any substantial computation.

Each bar in Figure 4 shows the time to complete the PostMark benchmark. The y-axis is logarithmic because of the substantial slowdown of synchronous I/O. The first bar shows results when ext3 is mounted asynchronously. As expected, this offers the best performance, since the file system buffers data in memory up to five seconds before writing it to disk. The second bar shows results using xsyncfs. Despite the I/O intensive nature of PostMark, the performance of xsyncfs is within 7% of the performance of ext3 mounted asynchronously. After examining the performance of xsyncfs, we determined that the overhead of tracking causal dependencies in the kernel accounts for most of the difference.

The third bar shows performance when ext3 is mounted synchronously. In this configuration, the writing process is blocked until its modifications are committed to the drive cache. Ext3 in synchronous mode is more than an order of magnitude slower than xsyncfs, even though xsyncfs provides stronger durability guarantees. Throughput is limited by the size of the drive cache; once the cache fills, subsequent writes block until some data in the cache is written to the disk platter.

The last bar in Figure 4 shows the time to complete the benchmark when ext3 is mounted synchronously and write barriers are used to prevent data loss when a power failure occurs. Since write barriers synchronously flush the drive cache twice for each file system transaction, ext3's performance is more than two orders of magnitude slower than that of xsyncfs.

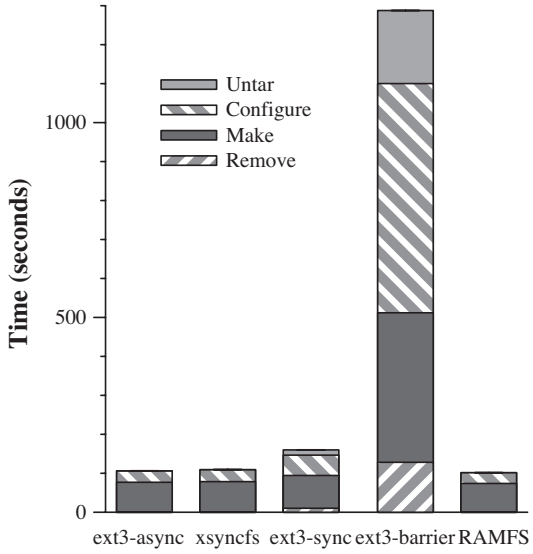


Fig. 5. The Apache build benchmark. This figure shows the time to run the Apache build benchmark. Each value is the mean of 5 trials—the relatively small error bars are 90% confidence intervals.

Due to the high cost of durability, high end storage systems sometimes use specialized hardware such as a nonvolatile cache to improve performance [Hitz et al. 1994]. This eliminates the need for write barriers. However, even with specialized hardware, we expect that the performance of ext3 mounted synchronously would be no better than the third bar in Figure 4, which writes data to a volatile cache. Thus, use of xsyncfs should still lead to substantial performance improvements for synchronous operations, even when the hard drive has a non-volatile cache of the same size as the volatile cache on our drive.

4.4 The Apache Build Benchmark

We next run a benchmark in which we untar the Apache 2.0.48 source tree into a file system, run `configure` in an object directory within that file system, run `make` in the object directory, and remove all files. The Apache build benchmark requires the file system to balance throughput and latency; it displays large amounts of screen output interleaved with disk I/O and computation.

Figure 5 shows the total amount of time to run the benchmark, with shadings within each bar showing the time for each stage. Comparing the first two bars in the graph, xsyncfs performs within 3% of ext3 mounted asynchronously. Since xsyncfs releases output as soon as the data on which it depends commits, output appears promptly during the execution of the benchmark.

For comparison, the bar at the far right of the graph shows the time to execute the benchmark using a memory-only file system, RAMFS. This provides a lower bound on the performance of a local file system, and it isolates the computation requirements of the benchmark. Removing disk I/O by running the

benchmark in RAMFS improves performance by only 8% over xsyncfs because the remainder of the benchmark is dominated by computation.

The third bar in Figure 5 shows that ext3 mounted in synchronous mode is 46% slower than xsyncfs. Since computation dominates I/O in this benchmark, any difference in I/O performance is a smaller part of overall performance. The fourth bar shows that ext3 mounted synchronously with write barriers is over 11 times slower than xsyncfs. If we isolate the cost of I/O by subtracting the cost of computation (calculated using the RAMFS result), ext3 mounted synchronously is 7.5 times slower than xsyncfs, and ext3 mounted synchronously with write barriers is over two orders of magnitude slower than xsyncfs. These isolated results are similar to the values that we saw for the PostMark experiments.

4.5 The MySQL Benchmark

We were curious to see how xsyncfs would perform with an application that implements its own group-commit strategy. We therefore ran a modified version of the OSDL TPC-C benchmark [OSDL 2006] using MySQL version 5.0.16 and the InnoDB storage engine. Since both MySQL and the TPC-C benchmark client are multi-threaded, this benchmark measures the efficacy of xsyncfs's support for shared memory. TPC-C measures the new order transactions per minute (NOTPM) a database can process for a given number of simultaneous client connections. The total number of transactions performed by TPC-C is approximately twice the number of new order transactions. TPC-C requires that a database provide ACID semantics, and MySQL requires either disabling the drive cache or using write barriers to provide durability. Therefore, we compare xsyncfs with ext3 mounted asynchronously using write barriers. We ran two versions of the benchmark. In the first version, the client ran on the same machine as the server; therefore, we modified that version of the benchmark to use UNIX sockets. This allows xsyncfs to propagate commit dependencies between the client and server on the same machine. To capture the benefits of propagating dependencies between the client and server, we ran a second version of the benchmark using two machines over a 100 Mbps Ethernet switch. In addition, we modified the benchmark to saturate the MySQL server by removing any wait times between transactions and creating a data set that fits completely in memory.

Figure 6(A) shows the NOTPM achieved as the number of clients is increased from 1 to 20. With a single client, MySQL completes three times as many NOTPM using xsyncfs. By propagating commit dependencies to both the MySQL server and the requesting client, xsyncfs can group-commit transactions from a single client, significantly improving performance. In contrast, MySQL cannot benefit from group-commit with a single client because it must conservatively commit each transaction before replying to the client.

When there are multiple clients, MySQL can group the commits of transactions from different clients. As the number of clients grows, the gap between xsyncfs and ext3 mounted asynchronously with write barriers shrinks. With 20 clients, xsyncfs improves TPC-C performance by 22%. When the number

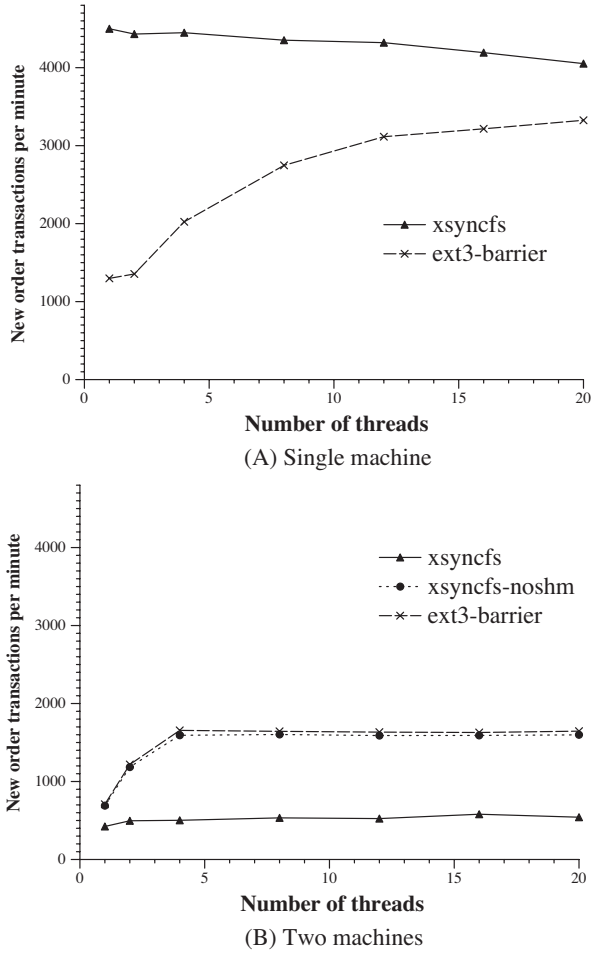


Fig. 6. The MySQL benchmarks. These figures show the new order transactions per minute when running a modified TPC-C benchmark on MySQL with varying numbers of clients. Figure (A) executes with both the client and server on a single machine. Figure (B) executes with the clients connected to the server over a 100 Mbps Ethernet switch. Each result is the mean of five trials—the error bars are 90% confidence intervals.

of clients reaches 32, the performance of ext3 mounted asynchronously with write barriers matches the performance of xsyncfs. From these results, we conclude that even applications, such as MySQL, that use a custom group-commit strategy, can benefit from external synchrony if the number of concurrent transactions is low to moderate.

Although ext3 mounted asynchronously without write barriers does not meet the durability requirements for TPC-C, we were still curious to see how its performance would compare to xsyncfs. With only one or two clients, MySQL executes 11% more NOTPM with xsyncfs than it executes with ext3 without write barriers. With four or more clients, the two configurations yield equivalent performance within experimental error.

Figure 6(B) shows the NOTPM achieved when the benchmark is run over a 100 Mbps Ethernet switch. Xsyncfs performs significantly more slowly than ext3 mounted asynchronously with write barriers. On each write to the file system, Xsyncfs conservatively propagates dependencies to every thread in the MySQL server. In addition, each outgoing network packet triggers a file system commit. We hypothesized that our conservative approach to handling shared memory created many false dependencies between threads, which degraded performance. To test our hypothesis, we modified Speculator so that it treated threads as independent processes; dependencies were not propagated to shared memory segments (we did retain all of the other avenues of propagation). The line labeled xsyncfs-noshm shows the NOTPM when xsyncfs does not propagate dependencies between threads; the performance of xsyncfs improves to nearly match that of ext mounted asynchronously with write barriers.

Two alternate strategies could be used to improve the performance of external synchrony with programs such as MySQL. First, if the threads are completing independent work, fine grained [Scales et al. 1996] or explicit [Hill et al. 1993] memory sharing could be used to transfer dependencies only when necessary. A second alternative would be to make the clients aware of external synchrony. By using speculative execution [Nightingale et al. 2006] we could expand the notion of what is external and propagate dependencies across the network.

4.6 The SPECweb99 Benchmark

Finally, we ran the SPECweb99 [Standard Performance Evaluation Corporation 2006] benchmark to examine the impact of external synchrony on a network-intensive, read-heavy application. In the SPECweb99 benchmark, multiple clients issue a mix of HTTP GET and POST requests. HTTP GET requests are issued for both static and dynamic content up to 1 MB in size. A single client, emulating 50 simultaneous connections, is connected to the server, which runs Apache 2.0.48, by a 100 Mb/s Ethernet switch. Since we use the default Apache settings, 50 connections are sufficient to saturate our server.

We felt that this benchmark would also be challenging for xsyncfs, since sending a network message externalizes state. Since xsyncfs only tracks causal dependencies on a single computer, it must buffer each message until the file system data on which that message depends has been committed. In addition to the normal log data written by Apache, the SPECweb99 benchmark writes a log record to the file system as a result of each HTTP POST. Thus, small file writes are common during benchmark execution—a typical 45 minute run has approximately 150,000 file system transactions.

As shown in Figure 7, SPECweb99 throughput using xsyncfs is within 8% of the throughput achieved when ext3 is mounted asynchronously. In contrast to ext3, xsyncfs guarantees that the data associated with each POST request is durable before a client receives the POST response. The third bar in Figure 7 shows that SPECweb99 using ext3 mounted synchronously achieves 6% higher throughput than xsyncfs. Unlike the previous benchmarks, SPECweb99 writes little data to disk, so most writes are buffered by the drive cache. The last

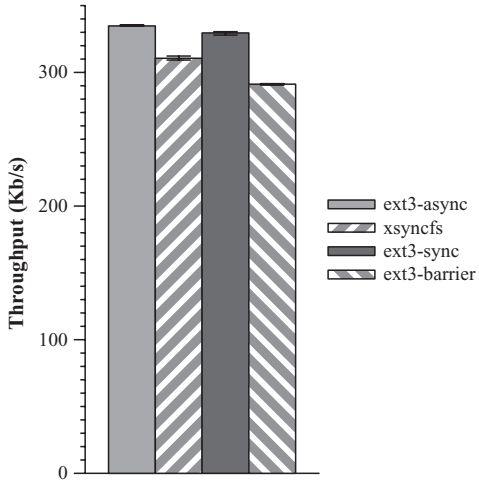


Fig. 7. Throughput in the SPECweb99 benchmark. This figure shows the mean throughput achieved when running the SPECweb99 benchmark with 50 simultaneous connections. Each result is the mean of three trials, with error bars showing the highest and lowest result.

Request size	ext3-async	xsyncfs	ext3-barrier
0–1 KB	0.064 (±0.025)	0.097 (±0.002)	0.122 (±0.001)
1–10 KB	0.150 (±0.034)	0.180 (±0.001)	0.206 (±0.001)
10–100 KB	1.084 (±0.052)	1.094 (±0.003)	1.114 (±0.005)
100–1000 KB	10.253 (±0.098)	10.072 (±0.066)	10.158 (±0.101)

Fig. 8. SPECweb99 latency results. This figure shows the mean time in seconds to request a file of a particular size during three trials of the SPECweb99 benchmark with 50 simultaneous connections. 90% confidence intervals are given in parentheses.

bar shows that xsyncfs achieves 7% better throughput than ext3 mounted synchronously with write barriers.

Figure 8 summarizes the average latency of individual HTTP requests. On average, xsyncfs adds no more than 33 ms of delay to each request when used instead of ext3 mounted asynchronously—this value is less than the commonly cited perception-threshold of 50 ms for human users [Flautner and Mudge 2002]. Thus, a user should perceive no difference in response time between xsyncfs and ext3 for HTTP requests. Although SPECweb99 does not require synchronous guarantees, we were curious to see how xsyncfs compared to ext3 mounted asynchronously with write barriers. The results are shown in the third column of Figure 8. For small files, xsyncfs adds less delay to each request than ext3 mounted synchronously with write barriers, while providing an equivalent guarantee to the user.

4.7 Benefit of Output-Triggered Commits

To measure the benefit of output-triggered commits, we also implemented an *eager commit* strategy for xsyncfs that triggers a commit whenever the file system is modified. The eager commit strategy still allows for group-commit,

Benchmark	Eager commits	Output-triggered commits	Speedup
PostMark (seconds)	9.879 (±0.056)	8.668 (±0.478)	14%
Apache (seconds)	111.41 (±0.32)	109.42 (±0.71)	2%
MySQL 1 client (NOTPM)	3323 (±60)	4498 (±73)	35%
MySQL 20 clients (NOTPM)	3646 (±217)	4052 (±200)	11%
SPECweb99 (Kb/s)	312 (±1)	311 (±2)	0%

Fig. 9. Benefit of output-triggered commits. This figure compares the performance of output-triggered commits with an eager commit strategy. Each result shows the mean of five trials, except SPECweb99, which is the mean of three trials. 90% confidence intervals are given in parentheses.

since multiple modifications are grouped into a single file system transaction while the previous transaction is committing. The next transaction will only start to commit once the commit of the previous transaction completes. The eager commit strategy attempts to minimize the latency of individual file system operations.

We executed the previous benchmarks using the eager commit strategy. Figure 9 compares results for the two strategies. The output-triggered commit strategy performs better than the eager commit strategy in every benchmark except SPECweb99, which creates so much output that the eager commit and output-triggered commit strategies perform very similarly. Since the eager commit strategy attempts to minimize the latency of a single operation, it sacrifices the opportunity to improve throughput. In contrast, the output-triggered commit strategy only minimizes latency after output has been generated that depends on a transaction; otherwise it maximizes throughput.

5. RELATED WORK

To the best of our knowledge, xsyncfs is the first local file system to provide high-performance synchronous I/O without requiring specialized hardware support or application modification. Further, xsyncfs is the first file system to use the causal relationship between file modifications and external output to decide when to commit data.

While xsyncfs takes a software-only approach to providing high-performance synchronous I/O, specialized hardware can achieve the same result. The Rio file cache [Chen et al. 1996] and the Conquest file system [Wang et al. 2002] use battery-backed main memory to make writes persistent. Durability is guaranteed only as long as the computer has power or the batteries remain charged.

Hitz et al. [1994] store file system journal modifications on a battery-backed RAM drive cache, while writing file system data to disk. We expect that synchronous operations on Hitz's hybrid system would perform no better than ext3 mounted synchronously without write barriers in our experiments. Thus xsyncfs could substantially improve the performance of such hybrid systems.

eNVy [Wu and Zwaenepoel 1994] is a file system that stores data on flash-based NVRAM. The designers of eNVy found that although reads from NVRAM were fast, writes were prohibitively slow. They used a battery-backed RAM write cache to achieve reasonable write performance. The write performance issues seen in eNVy are similar to those we experienced writing data to

commodity hard drives. Therefore it is likely that xsyncfs could also improve performance for flash file systems.

Xsyncfs's focus on providing both strong durability and reasonable performance contrasts sharply with the trend in commodity file systems toward relaxing durability to improve performance. Early file systems, such as FFS [McKusick et al. 1984] and the original UNIX file system [Ritchie and Thompson 1974], introduced the use of a main memory buffer cache to hold writes until they are asynchronously written to disk. Early file systems suffered from potential corruption when a computer lost power or an operating system crashed. Recovery often required a time consuming examination of the entire state of the file system (e.g., running `fsck`). For this reason, file systems such as Cedar [Hagmann 1987] added the complexity of a write-ahead log to enable fast, consistent recovery of file system state. Yet, as was shown in our evaluation, journaling data to a write-ahead log is insufficient to prevent file system corruption if the drive cache reorders block writes. An alternative to write-ahead logging, Soft Updates [Seltzer et al. 2000], carefully orders disk writes to provide consistent recovery. Xsyncfs builds on this prior work, since it writes data after returning control to the application, and uses a write-ahead log. Thus external synchrony could improve the performance of synchronous I/O with other journaling file systems, such as JFS [Best 2000] or ReiserFS [Namesys 2006].

Fault tolerance researchers have long defined consistent recovery in terms of the output seen by the outside world [Elnozahy et al. 2002; Lowell et al. 2000; Strom and Yemini 1985]. For example, the *output commit* problem requires that before a message is sent to the outside world, the state from which that message is sent must be preserved. In the same way, we argue that the guarantees provided by synchronous disk I/O should be defined by the output seen by the outside world, rather than by the results seen by local processes.

It is interesting to speculate on why the principle of outside observability is widely known and used in fault tolerance research, yet is new to the domain of general purpose applications and I/O. We believe this dichotomy arises from the different *scope* and *standard* of recovery in the two domains. In fault tolerance research, the scope of recovery is the entire process; hence not using the principle of outside observability would require a synchronous disk I/O at every change in process state. In general purpose applications, the scope of recovery is only the I/O issued by the application (which can be viewed as an application-specific recovery protocol). Hence it is feasible, though still slow, to issue each I/O synchronously. In addition, the standard for recovery in fault tolerance research is well defined: a recovery system should lose no visible output. In contrast, the standard for recovery in general purpose systems is looser: asynchronous I/O is common, and even synchronous I/O is usually committed synchronously only to the volatile hard drive cache.

Our implementation of external synchrony draws upon two other techniques from the fault tolerance literature. First, buffering output until the commit, is similar to deferring message sends until the commit [Lowell and Chen 1998]. Second, tracking causal dependencies to identify what and when to commit is similar to causal tracking in message logging protocols [Elnozahy and Zwaenepoel 1992]. We use these techniques in isolation, to improve

performance and maintain the appearance of synchronous I/O. We also use these techniques in combination via output-triggered commits, which automatically balance throughput and latency.

Transactions, provided by operating systems such as QuickSilver [Schmuck and Wylie 1991], TABS [Spector et al. 1985], and Locus [Weinstein et al. 1985], and by transactional file systems [Liskov and Rodrigues 2004; Paxton 1979], also give the strong durability and ordering guarantees that are provided by xsyncfs. In addition, transactions provide atomicity for a set of file system operations. However, transactional systems typically require that applications be modified to specify transaction boundaries. In contrast, use of xsyncfs requires no such modification.

6. CONCLUSION

It is challenging to develop simple and reliable software systems if the foundations upon which those systems are built are unreliable. Asynchronous I/O is a prime example of one such unreliable foundation. OS crashes and power failures can lead to loss of data, file system corruption, and out-of-order modifications. Nevertheless, current file systems present an asynchronous I/O interface by default, because the performance penalty of synchronous I/O is assumed to be too large.

In this article, we have proposed a new abstraction, external synchrony, that preserves the simplicity and reliability of a synchronous I/O interface, yet performs approximately as well as an asynchronous I/O interface. Based on these results, we believe that externally synchronous file systems such as xsyncfs can provide a better foundation for the construction of reliable software systems.

ACKNOWLEDGMENTS

We thank Manish Anand, Evan Cooke, Anthony Nicholson, Dan Peek, Sushant Sinha, Ya-Yunn Su, Rob Pike, George Candea, Greg Ganger, and the anonymous reviewers for feedback on this article.

REFERENCES

BEST, S. 2000. JFS overview. Tech. Rep., IBM, http://www-128.ibm.com/developerworks/linux/library/l-jfs.html.

CHEN, P. M., NG, W. T., CHANDRA, S., AYCOCK, C., RAJAMANI, G., AND LOWELL, D. 1996. The Rio file cache: Surviving operating system crashes. In *Proceedings of the 7th International Conference on Architectural Support for Programming Languages and Operating Systems*. Cambridge, MA, 74–83.

ELNOZAHY, E. N., ALVISI, L., WANG, Y.-M., AND JOHNSON, D. B. 2002. A survey of rollback-recovery protocols in message-passing systems. *ACM Comput. Surv. 34*, 3, 375–408.

ELNOZAHY, E. N. AND ZWAENEPOEL, W. 1992. Manetho: transparent rollback-recovery with low overhead, limited rollback, and fast output commit. *IEEE Trans. Comput. 41,* 5, 526–531.

FLAUTNER, K. AND MUDGE, T. 2002. Vertigo: automatic performance-setting for Linux. In *Proceedings of the 5th Symposium on Operating Systems Design and Implementation*. Boston, MA, 105–116.

HAGMANN, R. 1987. Reimplementing the Cedar file system using logging and group commit. In *Proceedings of the 11th ACM Symposium on Operating Systems Principles*. Austin, TX, 155–162.

HILL, M. D., LARUS, J. R., REINHARDT, S. K., AND WOOD, D. A. 1993. Cooperative shared memory: software and hardware for scalable multiprocessors. *ACM Trans. Comput. Syst. 11,* 4, 300–318.

HITZ, D., LAU, J., AND MALCOLM, M. 1994. File system design for an NFS file server appliance. In *Proceedings of the Winter USENIX Technical Conference*.

KATCHER, J. 1997. PostMark: A new file system benchmark. Tech. rep. TR3022, Network Appliance.

LAMPORT, L. 1978. Time, clocks, and the ordering of events in a distributed system. *ACM Commun. 21,* 7, 558–565.

LISKOV, B. AND RODRIGUES, R. 2004. Transactional file systems can be fast. In *Proceedings of the 11th SIGOPS European Workshop*. Leuven, Belgium.

LOWELL, D. E., CHANDRA, S., AND CHEN, P. M. 2000. Exploring failure transparency and the limits of generic recovery. In *Proceedings of the 4th Symposium on Operating Systems Design and Implementation*. San Diego, CA.

LOWELL, D. E. AND CHEN, P. M. 1998. Persistent messages in local transactions. In *Proceedings of the 1998 Symposium on Principles of Distributed Computing*. 219–226.

MCKUSICK, M. K. 2006. Disks from the perspective of a file system. *;login: 31,* 3, 18–19.

MCKUSICK, M. K., JOY, W. N., LEFFLER, S. J., AND FABRY, R. S. 1984. A fast file system for unix. *ACM Trans. Comput. Syst. 2,* 3, 181–197.

MySQL AB. 2006. *MySQL Reference Manual*. MySQL AB. http://dev.mysql.com/.

NAMESYS. 2006. *ReiserFS*. Namesys, http://www.namesys.com/.

NIGHTINGALE, E. B., CHEN, P. M., AND FLINN, J. 2006. Speculative execution in a distributed file system. *ACM Trans. Comput. Syst. 24,* 4, 361–392.

OSDL 2006. *OSDL Database test 2*. OSDL, http://www.osdl.org/.

PAXTON, W. H. 1979. A client-based transaction system to maintain data integrity. In *Proceedings of the 7th ACM Symposium on Operating Systems Principles*. 18–23.

PRABHAKARAN, V., BAIRAVASUNDARAM, L. N., AGRAWAL, N., GUNAWI, H. S., ARPACI-DUSSEAU, A. C., AND ARPACI-DUSSEAU, R. H. 2005. IRON file systems. In *Proceedings of the 20th ACM Symposium on Operating Systems Principles*. Brighton, UK, 206–220.

QIN, F., TUCEK, J., SUNDARESAN, J., AND ZHOU, Y. 2005. Rx: treating bugs as allergies—a safe method to survive software failures. In *Proceedings of the 20th ACM Symposium on Operating Systems Principles*. Brighton, UK, 235–248.

RITCHIE, D. M. AND THOMPSON, K. 1974. The UNIX time-sharing system. *ACM Commun. 17,* 7, 365–375.

SCALES, D. J., GHARACHORLOO, K., AND THEKKATH, C. A. 1996. Shasta: a low overhead, software-only approach for supporting fine-grain shared memory. In *Proceedings of the 7th Symposium on Architectural Support for Programming Languages and Operating Systems (ASPLOSVII)*. 174–185.

SCHMUCK, F. AND WYLIE, J. 1991. Experience with transactions in QuickSilver. In *Proceedings of the 13th ACM Symposium on Operating Systems Principles*. 239–253.

SELTZER, M. I., GANGER, G. R., MCKUSICK, M. K., SMITH, K. A., SOULES, C. A. N., AND STEIN, C. A. 2000. Journaling versus soft updates: asynchronous meta-data protection in file systems. In *Proceedings of the USENIX Annual Technical Conference*. San Diego, CA, 18–23.

SILBERSCHATZ, A. AND GALVIN, P. B. 1998. *Operating System Concepts*, 5th ed. Addison Wesley. 27.

SLASHDOT. 2005. *Your hard drive lies to you*. Slashdot. http://hardware.slashdot.org/article.pl?sid=05/05/13/0529252.

SPECTOR, A. Z., DANIELS, D., DUCHAMP, D., EPPINGER, J. L., AND PAUSCH, R. 1985. Distributed transactions for reliable systems. In *Proceedings of the 10th ACM Symposium on Operating Systems Principles*. Orcas Island, WA, 127–146.

STANDARD PERFORMANCE EVALUATION CORPORATION. 2006. SPECweb99. Standard Performance Evaluation Corporation, http://www.spec.org/web99.

STROM, R. E. AND YEMINI, S. 1985. Optimistic recovery in distributed systems. *ACM Trans. Comput. Syst. 3,* 3, 204–226.

WANG, A.-I. A., REIHER, P., POPEK, G. J., AND KUENNING, G. H. 2002. Conquest: better performance through a disk/persistent-RAM hybrid file system. In *Proceedings of the USENIX Annual Technical Conference*. Monterey, CA.

WEINSTEIN, M. J., THOMAS W. PAGE, J., LIVEZEY, B. K., AND POPEK, G. J. 1985. Transactions and synchronization in a distributed operating system. In *Proceedings of the 10th ACM Symposium on Operating Systems Principles*. Oreas Island, WA, 115–126.

WU, M. AND ZWAENEPOEL, W. 1994. eNVy: a non-volatile, main memory storage system. In *Proceedings of the 6th International Conference on Architectural Support for Programming Languages and Operating Systems*. San Jose, CA, 86–97.

Received September 2007; revised June 2008; accepted June 2008