BRAVO – Biased Locking for Reader-Writer Locks

Dave Dice
Oracle Labs
dave.dice@oracle.com

Alex Kogan
Oracle Labs
alex.kogan@oracle.com

Abstract
Designers of modern reader-writer locks confront a difficult trade-off related to reader scalability. Locks that have a compact memory representation for active readers will typically suffer under high intensity read-dominated workloads when the “reader indicator” state is updated frequently by a diverse set of threads, causing cache invalidation and coherence traffic. Other designs, such as cohort reader-writer locks[4], use distributed reader indicators, one per NUMA node. This improves reader-reader scalability, but also increases the size of each lock instance.

We propose a simple transformation, BRAVO, that augments any existing reader-writer lock, adding just two integer fields to the lock instance. Readers make their presence known to writers by hashing their thread’s identity with the lock address, forming an index into a visible readers table. Readers attempt to install the lock address into that element in the table, making their existence known to potential writers. All locks and threads in an address space can share the visible readers table. Updates by readers tend to be diffused over the table, resulting in a NUMA-friendly design. Crucially, readers of the same lock tend to write to different locations in the array, reducing coherence traffic.

Specifically, BRAVO allows a simple compact lock to be augmented so as to provide scalable concurrent reading but with only a modest increase in footprint.

CCS Concepts  Software and its engineering → Multithreading; Mutual exclusion; Concurrency control; Process synchronization;

Keywords Reader-Writer Locks, Synchronization, Concurrency Control

1 Introduction
1.1 The BRAVO Algorithm
BRAVO transforms any existing reader-writer lock $A$ into BRAVO-$A$, which provides scalable reader acquisition. We say $A$ is the underlying lock in BRAVO-$A$. In typical circumstances $A$ might be a simple compact lock that suffers under high levels of reader concurrency. BRAVO-$A$ will also be compact, but is NUMA-friendly and offers scalability in the presence of frequently arriving concurrent readers.

BRAVO extends A’s structure with a new $\text{RBias}$ boolean field. Arriving readers first check the $\text{RBias}$ field, and, if found set, then hash the address of the lock with a value reflecting the calling thread’s identity to form an index into the visible readers table. (This table is shared by all locks and threads in an address space. In all our experiments we sized the table at 1024 entries. Each array element, or slot, is a either null or a pointer to a reader-writer lock instance). The reader then uses an atomic compare-and-swap “CAS” operator to attempt to change the element at that index from null to the address of the lock, publishing its existence to potential writers. If the CAS is successful then the reader rechecks the $\text{RBias}$ field to ensure it remains set. If so, the reader has successfully gained read permission and can enter the critical section. Upon completing the critical section the reader executes the complementary operation to release read permission, simply storing null into that slot. We refer to this as the fast-path.

If the recheck operation above happens to fail, as would be the case if a writer intervened and cleared $\text{RBias}$ and the reader lost the race, then the reader simply clears the slot and reverts to the traditional slow-path where it acquires read permission via the underlying lock. Similarly, if the initial check of $\text{RBias}$ found the flag clear, or the CAS failed because of collisions in the array – the slot was found to be populated – then control diverts to the traditional slow-path. After a slow-path reader acquires read permission from the underlying lock, it enters and executes the critical section, and then at unlock time releases read permission via the underlying lock.

Active readers can make their existence public in two ways: either via the visible readers table (fast-path), or via the traditional underlying reader-writer lock (slow-path). Our mechanism allows both slow-path and fast-path readers simultaneously. Absent hash collisions, concurrent fast-path readers will write to different locations in the visible readers table. Collisions are benign, and impact performance but not correctness.

Arriving writers first acquire write permission on the underlying reader-writer lock. Having done so, they then check the $\text{RBias}$ flag. If set, the writer must perform revocation, first clearing the $\text{RBias}$ flag and then scanning all the elements of the visible readers table checking for conflicting fast-path readers. If any elements match the lock, the writer must wait for that fast-path reader to depart and clear the slot. If lock $L$ has 2 fast-path active readers, for instance, then $L$ will appear twice in the array. Scanning the array might appear to be onerous, but in practice the sequential scan is assisted by the automatic hardware prefetchers present in modern CPUs. We observe an amortized scan rate of about 1.1 nanoseconds...
per element on our system-under-test (described later). Having checked $R\text{Bias}$ and performed revocation if necessary, the writer then enters the critical section. At unlock-time, the writer simply releases write permission on the underlying reader-writer lock. The only difference for writers under BRAVO is the requirement to check and potentially revoke reader bias if $R\text{Bias}$ was found set. Revocation is only required on transitions from reading to writing and when $R\text{Bias}$ was previously set. Note that writers only scan the reader visibility array, and never write into it.

BRAVO provides a dual existence representation for active readers, with their existence reflected in either the array or the underlying lock. Writers resolve read-vs-write conflicts against fast-path readers via the visible readers table and against slow-path readers via the underlying reader-writer lock.

In our first prototypes we set $R\text{Bias}$ in the reader slow-path based on a low-cost Bernoulli trial with probability $P = 1/100$ using a thread-local Marsaglia XOR-Shift [21] pseudorandom number generator. For safety, readers can only set $R\text{Bias}$ while they hold read permission on the underlying reader-writer lock, avoiding interactions with writers.

While this simplistic policy for enabling bias worked well in practice, we were concerned about situations where we might have enabled too eagerly, and incur frequent revocation to the point where BRAVO-A might be slower than A. Specifically, the worst-case scenario would be where slow readers repeatedly set $R\text{Bias}$, only to have it revoked immediately by a writer. The key additional cost in BRAVO is the revocation step, which executes under the underlying write lock and thus serializes operations associated with the lock.

As such, we measure the latency of revocation and multiply that period by $N$, a configurable parameter, and then inhibit the subsequent setting of bias in the reader slow-path for that period, bounded the worst-case expected slow-down from BRAVO for writers to $1/(N + 1)$. Our specific performance goal is *primum non nocere* – first, do no harm, with BRAVO-A never underperforming A by any significant margin on any workload. This tactic is simple and effective, but excessively conservative, taking into account only the worst-case performance penalty imposed by BRAVO, and not accounting for any potential benefit conferred by the BRAVO fast-path. Furthermore, measuring the revocation duration also incorporates the waiting time, as well as the scanning time, yielding a conservative over-estimate of the revocation scan cost and resulting in less aggressive use of reader bias. Despite these concerns, we find this policy yields good and predictable performance. For all benchmarks in this paper we used $N = 9$ yielding a worst-case writer slow-down bound of about 10%. Our policy required adding a second BRAVO-specific timestamp field $\text{InhibitUntil}$ which reflects the earliest time at which slow readers should reenable bias.

We note that if the underlying lock algorithm A has reader preference or writer preference, then BRAVO-A will exhibit that same property.

BRAVO acts as an accelerator layer, as readers can always fall back to the traditional underlying lock to gain read access. The benefit arises from avoiding coherence traffic on the centralized reader indicators in the underlying lock, and instead relying on updates to be diffused over the visible readers table. Fast-path readers write only into the visible readers table, and not the lock instance proper.

Write performance, and the scalability of read-vs-write and write-vs-write behavior depends solely on the underlying lock. Under high write intensity, with write-vs-write and write-vs-read conflicts, the performance of BRAVO devolves to that of the underlying lock. BRAVO accelerates reads only.

Listing-1 depicts a pseudo-code implementation of the BRAVO algorithm.

## 2 Related Work

**Reader-Indicator Design** Readers that are active – currently executing in a reader critical section – must be visible to potential writers. Writers must be able to detect active readers in order to resolve read-vs-write conflicts, and wait for active readers to depart. The mechanism through which readers make themselves visible is the *reader indicator*. Myriad designs have been described in the literature. At one end of the spectrum we find a centralized reader indicator with a integer field within each reader-writer lock instance that reflects the number of active readers. Readers use atomic instructions to increment and decrement this field. Classic example of such locks are [28] [22]. Another reader-writer lock algorithm having a compact centralized reader indicator is Brandenburg and Anderson’s Phase-Fair Ticket lock, designated PF-T in [3], where the reader indicator is encoded in two central fields. We refer to this algorithm as “BA” throughout the remainder of this paper. Such approaches are compact, having a small per-lock footprint, and simple, but, because of coherence traffic, do not scale in the presence of concurrent readers that are arriving and departing frequently [9, 10].

To address this concern, many designs turn toward distributed reader indicators. Each cohort reader-writer lock [4], for instance, uses a per-NUMA node reader indicator.

---

1. Addition costs associated with BRAVO include futile atomic operations from collisions, and sharing or false-sharing arising from near-collisions in the table. Our simplified cost model ignores these secondary factors. We note that the odds of collision are equivalent to those given by the “Birthday Paradox” [33] and that the general problem of deciding to set bias is equivalent to the classic “ski-rental” problem [32].
2. Our approach conservatively forgoes the potential of better performance afforded by the aggressive use of reader bias in order to limit the possibility of worsened performance[31].
3. We observe that it is trivial to collapse $R\text{Bias}$ and $\text{InhibitUntil}$ into just a single field. For clarity, we did not do so in our implementation.
4. Some implementations protect the reader indicator with a central lock.
While distributed reader indicators improve scalability, they also significantly increase the footprint of a lock instance, with each reader indicator residing on its own private cache line or sector. (On Intel we use 128 bytes). In addition, the size of the lock is variable with the number of nodes, and not known at compile-time, precluding simple static preallocation of locks. Writers are also burdened with the overhead of checking multiple reader indicators.

At the extreme other end of the spectrum we find designs with reader indicators assigned per-CPU or per-thread [1, 6, 16, 30]. These designs promote read-read scaling but have a large variable-sized footprint. They also favor readers in that writers must traverse and examine all the reader-indicators to resolve read-vs-write conflicts, possibly imposing a performance burden on writers. We note there are a number of such distributed locks: a set of reader-indicators coupled with a central mutual exclusion lock for writer permission, as found in cohort locks; sets of mutexes where readers must acquire one mutex and writers must acquire all mutexes; or sets of reader-writer locks where readers must acquire read permission on one lock, and writers must acquire write permission on all locks.

To reduce the impact on writers, some designs use a tree of distributed counters [19] where the root element contains a sum of the indicators with the subtrees.

In addition to distributing or dispersing the counters, individual counters can themselves be further split into constituent ingress and egress fields to further reduce write sharing. Arriving readers increment the ingress field and departing readers increment the egress field. Cohort reader-writer locks use this approach.

BRAVO takes a different approach, opportunistically representing active readers in the shared global reader visibility table. The array is fixed in size and shared over all threads and locks within an address space. Each BRAVO lock has, in addition to the underlying reader-writer lock, a boolean flag that indicates if reader bias is currently enabled for that lock. Publication of active readers in the array is strictly optional and best-effort. A reader can always fall back to acquiring read permission via the underlying reader-writer lock. BRAVO is NUMA-friendly, but unlike most other NUMA-aware reader-writer locks, it does not need to understand or otherwise query the system topology, further simplifying the design and reducing dependencies.

**Optimistic Invisible Readers** Synchronization constructs such as seqlocks [5, 14, 18] allow concurrent readers, but forgo the need for readers to make themselves visible. Critics, readers don’t write to synchronization data and thus do not induce coherence traffic. Instead, writers update state – typically a modification counter – to indicate that updates have occurred. Readers check that counter at the start and then again at the end of their critical section, and if writers were active or the counter changed, the readers

---

```python
class BRAVOLock<T>:
    int RBias
    Time InhibitUntil
    T Underlying

    ## Shared global :
    RWLock * VisibleReaders [1024]
    int N = 9     # slow-down guard

    def Reader(BRAVOLock * L):
        BRAVOLock * slot = null
        if L.RBias :
            slot = VisibleReaders + Hash(L, Self)
            if CAS(slot, null, L) == null :
                # CAS succeeded
                # store-load fence required on TSO
                # typically subsumed by CAS
                if L.RBias :
                    # recheck
                    goto EnterCS   # fast path
                *slot = null    # raced - RBias changed
                slot = null
        # Slow path
        assert slot == null
        AcquireRead (L.Underlying)
        if L.RBias == 0 and Time() >= L.InhibitUntil :
            L.RBias = 1
        EnterCS:
        ReaderCriticalSection()
        if slot != null :
            assert *slot == L
            *slot = null
        else :
            ReleaseRead (L.Underlying)

    def Writer(BRAVOLock * L):
        AcquireWrite (L.Underlying)
        if L.RBias :
            # revoke bias
            # store-load fence required on TSO
            L.RBias = 0
            auto start = Time()
            for i in xrange(VisibleReaders):
                while VisibleReaders[i] == L :
                    Pause()
                auto now = Time()
                # primum non-nocere :
                # limit and bound slow-down
                # arising from revocation overheads
                L.InhibitUntil = now + ((now - start) * N)
        WriterCriticalSection()
        ReleaseWrite (L.Underlying)
```

Listing 1. Simplified Python-like Implementation of BRAVO

---

While BRAVO is topology oblivious, it does require high-resolution low-latency means of reading the system clock. We further expect that reading the clock is scalable, and that concurrent readers do not interfere with each other. On systems with modern Intel CPUs and Linux kernels the RDTSCP instruction or clock_gettime(CLOCK_MONOTONIC) fast system call suffices.
self-abort and retry. An additional challenge for seqlocks is that readers can observe inconsistent state, and special care must be taken to constrain the effects and avoid errant behavior in readers. Often, non-trivial reader critical sections must be modified to safely tolerate optimistic execution. The StampedLock[23] facility in java.util.concurrent consists of a reader writer lock coupled with a seqlock, providing 3 modes: classic pessimistic write locking, classic pessimistic read locking, and optimistic reading.

To avoid the problem where optimistic readers might see inconsistent state, transactional lock elision [11, 15, 17, 25] based on hardware transactional memory can be used. Readers are invisible and do not write to shared data. Such approaches can be helpful, but are still vulnerable to indefinite abort and progress failure. In addition, the hardware transactional memory facilities required to support lock elision are not available on all systems, and are usually best-effort, without any guaranteed progress, requiring some type of fallback to pessimistic mechanisms.

**Biased Locking** BRAVO draws inspiration from biased locking[8, 13, 24, 27, 29]. Briefly, biased locking allows the same thread to repeatedly acquire and release a lock without requiring atomic instructions, except on the initial acquisition. If another thread attempted to acquire the lock, then expensive revocation was required to wrest bias from the original thread. The lock would then revert to normal non-biased mode for some period before again becoming potentially eligible for bias. (Conceptually, we can think of the lock as just being left in the locked state until there is contention. Subsequent lock and unlock operations by the original thread are ignored – the unlock operation is deferred until contention arises). Biased locking was a response to the CPU-local latencies incurred by atomic instructions on early Intel and SPARC processors and to the fact that locks in Java were often dominated by a single thread. Subsequently, processor designers have addressed the latency concern, rendering biased locking less profitable.

BRAVO is suitable for read-dominated workloads, allowing a fast-path for readers when reader bias is enabled for a lock. If a write request is issued against a reader-biased lock, reader bias is disabled and and revocation (scanning of the visible readers table) is required, shifting some cost from readers to writers. Classic biased locking provides benefit by reducing the number of atomic operations and improving latency. It does not improve scalability. BRAVO reader-bias, however, can improve both latency and scalability by reducing coherence traffic on the reader indicators in the underlying reader-writer lock.

Classic biased locking identifies a preferred thread, while BRAVO identifies a preferred access mode. That is, BRAVO biases toward a mode instead of thread identity.

### 3 Empirical Evaluation

All data was collected on an Oracle X5-2 system. The system has 2 sockets, each populated with an Intel Xeon E5-2699 v3 CPU running at 2.30GHz. Each socket has 18 cores, and each core is 2-way hyperthreaded, yielding 72 logical CPUs in total. The system was running Ubuntu 18.04 with a stock Linux version 4.15 kernel, and all software was compiled using the provided GCC version 7.3 toolchain at optimization level “-O3”. 64-bit code was used for all experiments. Factory-provided system defaults were used in all cases, and Turbo mode was left enabled. In all cases default free-range unbound threads were used.

We implemented all locks within LD_PRELOAD interposition libraries that expose the standard POSIX pthread_rwlock_t programming interface. The framework was made available by the authors of [4]. This allows us to change lock implementations by varying the LD_PRELOAD environment variable and without modifying the application code that uses reader-writer locks.

In the following figures “BA” refers to the Brandenburg-Anderson PF-T lock; “Cohort-RW” refers to the C-RW-WP lock from [4]; “Per-CPU” reflects a lock that consists of an array of BA locks, one for each CPU where readers acquire read-permission on the sub-lock associated with their CPU, and writers acquire writer permission on all the sub-locks; and “BRAVO-BA” reflects BRAVO implemented on top of BA. Except where otherwise noted, we plot the number of concurrent threads on the X-axis, and aggregate throughput on the Y-axis. We report the median of 5 independent runs for each data point.

We note that BRAVO works well when layered over the underlying Linux pthread_rwlock_t implementation. Furthermore, most such implementations are compact with central reader indicators. We opted to use BA as the underlying lock, however, as it has well-defined fairness properties.

We use a 128 byte sector size on Intel processors for alignment to avoid false sharing. The unit of coherence is 64 bytes throughout the cache hierarchy, but 128 bytes is required because of the adjacent cache line prefetch facility where pairs of lines are automatically fetched together. BA requires just 128 bytes – 4 32-bit integer fields with the overall size rounded up to the next sector boundary. BRAVO-BA adds the 8-byte InhibitUntil field, which contains a timestamp, and the 4-byte RBias field. Rounding up to the sector size, this still yields a 128 byte lock instance. Per-CPU consists of one instance of BA for each logical CPU, yielding a lock size of 926 bytes on our 72-way system. Cohort-RW consists of one reader indicator (128 bytes) per NUMA node, a central location for state (128 bytes) and a full cohort mutex [12] to provide writer exclusion. In turn, the cohort mutex requires one 128-byte sub-lock per NUMA node, and another sector for central state, for a total of 768 bytes. (While our implementation did not do so, we note that a more space
aggressive implementation of Cohort-RW could collocate the per-node reader indicators with the mutex sub-locks, and the central state for the read-write lock with its associated cohort mutex, yielding a size of 384 bytes). The size of BA and BRAVO-BA are fixed, and known at compile-time, while the size of Per-CPU varies with the number of logical CPUs, and the size of Cohort-RW varies with the number of NUMA nodes. Finally, we observe that BRAVO allows more relaxed approach toward the alignment and padding of the underlying lock. Since fast-path readers don’t mutate the underlying lock fields, the designer can reasonable forgo alignment and padding on that lock, without trading off reader scalability.

BRAVO also requires the visible readers table. With 1024 entries on a system with 64-bit pointers, the additional footprint is 8Kb. The table is aligned and sized to minimize the number of underlying pages (reducing TLB footprint) and to eliminate false sharing from variables that might be placed adjacent to the table.

3.1 Alternator

Figure 1 shows the results of our Alternator benchmark. The benchmark spawns the specified number of concurrent threads, which organize themselves into a logical ring, each waiting for notification from its “left” sibling. Notification is accomplished via setting a thread-specific variable and waiting is via simple busy-waiting. Once notified, the thread acquires and then immediately releases read permission on a shared reader-writer lock. Next the thread notifies its “right” sibling and then again itself waits. There are no writers, and there is no concurrency between readers. At most one reader is active at any given moment. At the end of a 10 second measurement interval the program reports the number of notifications. The BA lock suffers as the lines underlying the reader indicators “slosh” and migrate from cache to cache. In contrast BRAVO-BA readers touch different locations in the visible readers table as they acquire and release read permissions. BRAVO enables reader-bias early in the run, and it remains set for the duration of the measurement interval. All locks experience a significant performance drop between 1 and 2 threads due to the impact of coherent communication for notification. Crucially, we see that BRAVO-BA outperforms the underlying BA by a wide margin, and is competitive with the much larger Per-CPU lock. In addition, the performance of BA can be seen to degrade as we add threads, whereas the performance of BRAVO-BA remains stable.

Since the hash function that associates a read locking request with an index is deterministic, threads repeatedly locking and unlocking a specific lock will enjoy temporal locality and reuse in the visible readers table.

3.2 test_rwlock

We next report results test_rwlock benchmark described in [7]. The benchmark was designed to evaluate the performance and scalability of reader-writer locks against the RCU (Read-Copy Update) synchronization mechanism. We used the following command-line: `test_rwlock T 1 10 -c 10 -e 10 -d 1000`. The benchmark launches 1 fixed-role writer thread and T fixed-role reader threads for a 10 second measurement interval. The writer loops as follows: acquire a central pthread_rwlock_t instance; execute 10 units of work, which entails counting down a local variable; release writer permission; execute a non-critical section for 1000 work units. Readers loop acquiring the central lock for reading, executing 10 steps of work in the critical section, and then release the lock. (The benchmark has no facilities to allow a non-trivial critical section for readers). At the end of the measurement interval the benchmark reports the sum of iterations completed by all the threads. As we can see in Figure 2, BRAVO-BA significantly outperforms BA, and even the Cohort-RW lock at higher thread counts. Because of the workload is extremely read-dominated, the Per-CPU lock yields the best performance, albeit with a very large footprint and only because of the relatively low write rate.

3.3 RWBench

Using RWBench – modeled on a benchmark of the same name described in [4] – we evaluated the reader-write lock algorithms over a variety of read-write ratios, ranging from write-intensive in Figure 3a (9 out of every 10 operations are writes) to read-intensive in Figure 4c (1 out of every 10000 operations are writes), demonstrating that BRAVO inflicts no harm for write-intensive workloads, but improves performance for more read-dominated workloads. RWBench

obtained from https://github.com/urcu/userspace-rcu/blob/master/tests/benchmark/test_rwlock.c and modified slightly to allow a fixed measurement interval.
launches $T$ concurrent threads for a 10 second measurement interval. Each thread loops as follows: using a thread-local pseudo-random generator, decide to write with probability $P$ via a Bernoulli trial; writers acquire a central reader-write lock for write permission and then execute 10 steps of a thread-local C++ std::mt19937 random number generator and then release write permission, while readers do the same, but under read permission; execute a non-critical section of $N$ steps of the same random-number generator where $N$ is a random number uniformly distributed in $[0, 200]$ with average and median of 100. At the end of the measurement interval the benchmark reports the total number of top-level loops completed.

In Figure-3a we see poor scalability over all the locks by virtue of the highly serialized write-heavy nature of the workload. Per-CPU fairs poorly as writes, which are common, need to scan the array of per-CPU sub-locks. Cohort-RW provides some benefit, while BRAVO-BA tracks closely to BA, providing neither benefit nor harm. The same behavior plays out in Figure-3b ($P = 1/2$) and Figure-3c ($P = 1/10$), although in the latter we see some scaling from Cohort-RW. In Figure-4a ($P = 1/100$) we begin to see BRAVO-BA outperforming BA at higher thread counts. Figure-4b ($P = 1/1000$) and Figure-4c ($P = 1/10000$ – extremely read-dominated) are fairly similar, with BRAVO-BA yielding performance similar to that of Per-CPU, Cohort-RW yielding modest scalability, and BA yield flat performance as increase the thread count.

3.4 rocksdb readwhilewriting

We next explore performance sensitivity to reader-writer in the rocksdb database[26]. We observed high frequency reader traffic arising from calls in ::Get() to db/memtable.cc
GetLock() in the readwhilewriting benchmark. In Figure-5 we see the performance of BRAVO-BA tracks that of Per-CPU and always exceeds that of Cohort-RW and the underlying BA.

Figure 4. RWBench

(a) RWBench with 1% writes (1/100)

(b) RWBench with .1% writes (1/1000)

(c) RWBench with .01% writes (1/10000)

Figure 5. rocksdb readwhilewriting

3.5 rocksdb hash_table_bench

Rocksdb also provides a benchmark to stress the hash table used by their persistent cache. The benchmark implements a central shared hash table as a C++ std::unordered_map protected by a reader-writer lock. The cache is pre-populated before the measurement interval. At the end of the 50 second measurement interval the benchmark reports the aggregate operation rate – reads, erases, insertions – per millisecond. A single dedicated thread loops, erasing random elements, and another dedicated thread loops inserting new elements with a random key. Both erase and insertion operations require write access. The benchmark launches \( T \) reader threads, which loop, running lookups on randomly selected keys. We vary \( T \) on the X-axis. All the threads execute operations back-to-back without a delay between operations. The benchmark makes frequent use of malloc-free operations in the std::unordered_map. The default malloc allocator fails to fully scale in this environment and masks any benefit conferred by improved reader-writer locks, so we instead used the index-aware allocator from [2].

\footnote{We used rocksdb version 5.13.4 with the following command line: `db_bench --benchmarks=readwhilewriting --memtablerep=cuckoo --duration=100 --num=10000 --allow_concurrent_memtable_write=0 --inplace_update_num_locks=1 --stats_interval=10000000`}

\footnote{https://github.com/facebook/rocksdb/blob/master/utilities/persistent_cache/hash_table_bench.cc run with the following command-line: `hash_table_bench -nread_thread=T -nsec=50`
uses distributed read indicators, but without the footprint or near the that of the best modern reader-writer locks that underly the lock. BRAVO provides read-read performance at a high rate. The approach is simple, effective, and yields improved performance for read-dominated workloads compared to commonly used compact locks. The key trade-off inherent in the design is the benefit accrued by reads against the potential slow-down induced by revocation. Even in mixed workloads, we limit any slow-down stemming from revocation costs and bounding harm, making the decision to use BRAVO simple. BRAVO incurs a very small footprint increase per lock instance, and also adds a shared table of fixed size that can be used by all threads and locks. BRAVO’s key benefit arises from reducing coherence cost that would normally be incurred by locks having a central reader indicator. Write performance is left unchanged relative to the underlying lock. BRAVO provides read-read performance at or near the that of the best modern reader-writer locks that uses distributed read indicators, but without the footprint or complexity of such locks. BRAVO is implicitly NUMA-friendly.

Future directions We identify a number of future directions for our investigation into BRAVO-based designs:

- Dynamic sizing of the visible readers table based on collisions. Large tables will have reduced collision rates but larger scan revocation overheads.
- The reader fast-path currently probes just a single location and reverts to the slow-path after a collision. We plan on using a secondary hash to probe an alternative location.
- Accelerate the revocation scan operation via SIMD instructions such as AVX. The visible reader table is usually sparsely populated, making it amenable to such optimizations. Non-temporal non-polluting loads may also be helpful for the scan operation.
- As noted, our current policy to enable bias is conservative, and leaves untapped performance. We intend to explore more sophisticated adaptive policies based on recent behavior and to use a more faithful cost model.
- An interesting variation is to implement BRAVO on top of an underlying mutex instead of a reader-write lock.
- We currently use a hash function to map a thread’s identity and the lock address to an index in the array. There is no particular requirement that the function that associates a read request with an index be deterministic. We plan on exploring other functions, using time or random numbers to form indices.
- In our current implementation arriving readers are blocked while a revocation scan is in progress. This could be avoided by using adding a mutex to each BRAVO-enhanced lock. Arriving writers immediately acquire this mutex, which resolves all write-write conflicts. They then perform revocation, if necessary; acquire the underlying reader-vs-write lock with write permission; execute the writer critical section; and finally release both the mutex and the underlying reader-write lock. The underlying reader-writer lock resolves read-vs-write conflicts. We note that this general technique can be readily applied to other existing locks, such as linux’s br-lock family [20]. By applying this optimization we can mitigate revocation costs by allowing readers to flow through the reader slow-path while revocation is running. This also reduces variance for the latency of read operations.

4 Conclusion
BRAVO easily composes with existing locks, preserving desirable properties of those underlying locks, and yielding a composite lock with improved read-read scalability. We specifically target read-dominated workloads with multiple concurrent threads that acquire and release read permission at a high rate. The approach is simple, effective, and yields improved performance for read-dominated workloads compared to commonly used compact locks. The key trade-off inherent in the design is the benefit accrued by reads against the potential slow-down imposed by revocation. Even in mixed workloads, we limit any slow-down stemming from revocation costs and bounding harm, making the decision to use BRAVO simple. BRAVO incurs a very small footprint increase per lock instance, and also adds a shared table of fixed size that can be used by all threads and locks. BRAVO’s key benefit arises from reducing coherence cost that would normally be incurred by locks having a central reader indicator. Write performance is left unchanged relative to the underlying lock. BRAVO provides read-read performance at or near the that of the best modern reader-writer locks that uses distributed read indicators, but without the footprint or complexity of such locks. BRAVO is implicitly NUMA-friend.

Future directions We identify a number of future directions for our investigation into BRAVO-based designs:

- Dynamic sizing of the visible readers table based on collisions. Large tables will have reduced collision rates but larger scan revocation overheads.
- The reader fast-path currently probes just a single location and reverts to the slow-path after a collision. We plan on using a secondary hash to probe an alternative location.
- Accelerate the revocation scan operation via SIMD instructions such as AVX. The visible reader table is usually sparsely populated, making it amenable to such optimizations. Non-temporal non-polluting loads may also be helpful for the scan operation.
- As noted, our current policy to enable bias is conservative, and leaves untapped performance. We intend to explore more sophisticated adaptive policies based on recent behavior and to use a more faithful cost model.
- An interesting variation is to implement BRAVO on top of an underlying mutex instead of a reader-write lock.
- We currently use a hash function to map a thread’s identity and the lock address to an index in the array. There is no particular requirement that the function that associates a read request with an index be deterministic. We plan on exploring other functions, using time or random numbers to form indices.
- In our current implementation arriving readers are blocked while a revocation scan is in progress. This could be avoided by using adding a mutex to each BRAVO-enhanced lock. Arriving writers immediately acquire this mutex, which resolves all write-write conflicts. They then perform revocation, if necessary; acquire the underlying reader-vs-write lock with write permission; execute the writer critical section; and finally release both the mutex and the underlying reader-write lock. The underlying reader-writer lock resolves read-vs-write conflicts. We note that this general technique can be readily applied to other existing locks, such as linux’s br-lock family [20]. By applying this optimization we can mitigate revocation costs by allowing readers to flow through the reader slow-path while revocation is running. This also reduces variance for the latency of read operations.

Acknowledgments
We thank Shady Issa for useful discussions about revocation and the cost model.

References
[1] 2013. Distributed Cache-Line Counter Scalable RW-Lock. (2013). http://concurrencyfreaks.blogspot.com/2013/09/distributed-cache-line-counter-scalable.html
[2] Yehuda Afek, Dave Dice, and Adam Morrison. 2011. Cache Index-aware Memory Allocation. In Proceedings of the International Symposium on Memory Management (ISMM ’11). ACM. http://doi.acm.org/10.1145/1993478.1993486
[3] R. B. Brandenburg and J. H. Anderson. 2010. Spin-Based Reader-Writer Synchronization for Multiprocessor Real-Time Systems. In Real-Time Systems Journal. https://doi.org/10.1007/s11214-010-0036-6
[4] Irina Calciu, Dave Dice, Yossi Lev, Victor Luchangco, Virendra J. Marathe, and Nir Shavit. 2013. NUMA-aware Reader-writer Locks (PPoPP). ACM. http://doi.acm.org/10.1145/2442516.2442532
[5] Jonathan Corbet. 2003. Driver porting: mutual exclusion with seqlocks. (2003). lwn.net/Articles/22818/
[6] Jonathan Corbet. 2010. Big reader locks. https://lwn.net/Articles/378911. (2010). Accessed: 2018-04-20.
