NVRAM - response recover linearzability *

[Extended Abstract]

Hagit Attiya
Department of
Computer-Science
Technion
Haifa, Israel
hagit@cs.technion.ac.il

Ohad Ben-Baruch
Department of
Computer-Science
Ben-Gurion University of the
Negev
Be'er Sheva, Israel
ohadben@post.bgu.ac.il

Danny Hendler
Department of
Computer-Science
Ben-Gurion University of the
Negev
Be'er Sheva, Israel
hendlerd@cs.bgu.ac.il

ABSTRACT

Categories and Subject Descriptors

D.1.3 [**Programming Techniques**]: Concurrent Programming—*Parallel programming*; D.2.8 [**Analysis of Algorithms and Problem Complexity**]: Tradeoffs between Complexity Measures; F.1.3 [**Complexity Measures and Classes**]: Relations among complexity measures

General Terms

Theory, Performance

Keywords

Mutual exclusion, shared-memory, lower bounds, total store ordering, time complexity, remote memory reference (RMR)

1. INTRODUCTION

Shared-memory multiprocessors are now prevalent anywhere from high-end server machines, through desktop and laptop computers to smartphones, accelerating the shift to concurrent, multi-threaded software. Concurrent software involves a collection of threads, each running a separate piece of code, possibly on a different core. Since threads may be delayed because of a variety of reasons (such as cache-misses, interrupts, page-faults, and scheduler preemption), shared-memory multiprocessors are *asynchronous* in nature.

Asynchrony is also related to reliability, since asynchronous algorithms that provide nonblocking progress properties (e.g., lock-freedom and wait-freedom) in an asynchronous environment with reliable processes continue to provide the same progress properties when *crash failures* are introduced. Informally speaking, this property holds because a process that crashes permanently at an arbitrary point in the execution of its algorithm is indistinguishable to the other processes from one that is merely very slow. Owing to its simplicity and intimate relationship with asynchrony, the crash-failure model is almost ubiquitous in the treatment of non-blocking shared memory algorithms.

The attention to the crash-failure model has so far mostly neglected the *crash-recovery* model, in which a failed process may be resurrected after it crashes. The reason is that the crash-recovery model is poorly matched to multi-core architectures with volatile SRAM-based caches and DRAM-based main memories: Any state stored in main memory is lost entirely in the event of a system crash or power loss, and recording recovery information in non-volatile secondary storage (e.g., on a hard disk drive or solid state drive) imposes overheads that are unacceptable for performance-critical tasks, such as synchronizing threads inside the operating system kernel.

This separation between volatile main memory and non-volatile secondary storage has led to a partitioning of the program state into operational data stored using in-memory data structures, and recovery data stored using sequential on-disk structures such as transaction logs. In the event of a system-wide failure, such as a power outage, in-memory data structures are lost and must be reconstructed entirely from the recovery data, making the software system that relies on them temporarily unavailable. As a result, the design of in-memory data structures emphasizes disposable constructs optimized for parallelism, in contrast to the structures that hold recovery data, which cannot be discarded upon a failure and which benefit less from parallelism since their performance is limited by the secondary storage.

Recent developments in non-volatile main memory (NVRAM) media foreshadow the eventual convergence of primary and secondary storage into a single layer in the memory hierarchy, combining the performance benefits of conventional main memory with the durability of secondary storage. Traditional log-based recovery techniques can be applied correctly in such systems but fail to take full advantage of the parallelism enabled by allowing processing cores to access recovery data directly using memory operations rather than slow block transfers from secondary storage.

The proposed research will investigate how the performance benefits of NVRAM can be harnessed for improv-

ing the robustness of shared-memory programs by allowing concurrent algorithms to efficiently recover from crash failures. This challenge requires a careful rethinking of recovery mechanisms and involves addressing both foundational and algorithmic research questions. Our emphasis would be on direct, non-transactional, approaches of implementing algorithms for *recoverable concurrent* (also called *persistent*) objects. Our work will target the following more specific goals:

1. Formulate an abstract model of NVRAM shared-memory multiprocessors, together with correctness and progress conditions, to facilitate rigorous analysis of the theoretic limitations of these systems, on the one hand, and construction of concurrent algorithms that support effective recovery from crash failures, on the other hand.

2. Investigate the construction of recoverable versions of primitive memory operations such as read, write, swap, compare-and-swap and fetch-and-add, deriving upper and lower bounds, and how these recoverable versions can be used to implement various applications.

3. Investigate the construction of recoverable versions of specific widely-used concurrent objects such as mutual-exclusion locks, stacks, queues and hash tables, deriving upper and lower bounds.

4. Provide a framework that incorporates additional aspects of shared-memory systems, like volatile caches, and methods for adapting results developed in the abstract model to real-world systems.

In the rest of this section, we provide required background. First, we briefly describe the standard shared-memory model. This is followed by a survey of correctness conditions that were proposed in recent years for defining the behavior of recoverable concurrent objects. We then survey previous work on transactional access to recoverable concurrent objects as well as more direct, non-transactional, implementations of such objects.

2. MODEL AND DEFINITIONS

2.1 Standard Shared-Memory Model

Base Objects and Primitive Operations.

Concurrent applications are programs that use *base objects*. In the simplest case, these are the shared memory locations provided by the multi-core system.

The system's hardware or operating system provide *primitive operations* that can be applied to memory locations. The simplest primitives, which are always assumed, are the familiar *read* and *write* operations, also called *load* and *store*. Modern architectures also provide stronger atomic *read-modify-write* primitives. The most notable of these is *compare-and-swap* (abbreviated *CAS*), which takes three arguments: a memory address, an old value, and a new value. If the address stores the old value, it is replaced with the new value; otherwise it is unchanged. The success or failure of this operation is then reported back to the program.

Another example of a widely-implemented strong primitive is *fetch-and-add*, which takes a memory address and an integer as its two arguments. When applied, it atomically adds the integer argument to the value stored at the memory address and returns the previous value.

Correctness Conditions.

When the application implements a concurrent data structure, it is necessary to specify its *semantics*, namely, the properties provided by values it returns, which determine the feasibility and complexity of implementing it. The *correctness* condition of a concurrent data structure specifies how to derive the semantics of concurrent implementations of the data structure from the corresponding *sequential specification* of the data structure. This requires to disambiguate the expected results of concurrent operations on the data structure.

Two common ways to do so are *sequential consistency* [18] and *linearizability* [14]. Both require that the values returned by the operations appear to have been returned by a sequential execution of the same operations; sequential consistency only requires this order to be consistent with the order in which each individual thread invokes the operations, while linearizability further requires this order to be consistent with the real-time order of non-overlapping operations. The standard shared-memory model assumes that shared memory is linearizable or at least sequentially consistent.

Progress Guarantees.

Progress (also called *liveness*) guarantees specify the conditions under which operations on a concurrent data-structure terminate. With *blocking* (lock-based) implementations, the failure of a single thread may halt the progress of all other threads, if the thread fails while holding a lock. Consequently, the progress of blocking implementations can be guaranteed only when every participating thread is *live*. (A thread is live if, whenever it begins executing an algorithm, it continues to take steps until the algorithm terminates.)

When locks are not used, threads should be able to make progress despite failures of other threads. There are several formal definitions capturing this intuitive requirement. The strongest liveness guarantee, *wait-freedom* [11], assures that every operation performed by a thread completes within a finite number of its own steps, regardless of the failures of other threads. The weaker *obstruction-freedom* guarantee [2, 13, 12] ensures only that if a thread executes in isolation (from some point on), without interleaved steps of other threads, then it will complete its operation within a finite number of steps.

2.2 Crash-Recovery Model

Linearizability is ill-equipped to specify the correctness criteria for implementations that support crash-recovery failures, since linearizability has no notion of an aborted or failed operation, and requires that a process finish one operation before it invokes the next. Extended versions of linearizability or, more generally, alternative definitions are required for specifying correctness for such implementations.

Aguilera and Frølund [1] proposed *strict linearizability* as a correctness condition for persistent concurrent objects, which treats the crash of a process as a response, either successful or unsuccessful, to the interrupted operation. A successful response means that the operation takes effect at some point between its invocation and the crash failure, and an unsuccessful response means that the operation does not

take effect at all. Strict linearizability preserves both *locality* [11] and program order. However, they proved that it precludes the implementation of some wait-free objects for certain machine models: there is no wait-free implementations of multi-reader single-writer (MRSW) registers from single-reader single-writer (SRSW) registers under strict linearizability.

Guerraoui and Levy [10] proposed *persistent atomicity*. It is similar to strict linearizability, but allows an operation interrupted by a failure to take effect before the subsequent invocation of the same process, possibly after the failure. They also proposed *transient atomicity*, which relaxes this criterion even further and allows an interrupted operation to take effect before the subsequent write response of the same process. Both conditions ensure that the state of an object will be consistent in the wake of a crash, but they do not provide locality: correct histories of separate objects, when merged, will not necessarily yield a correct composite history.

Berryhill et al. [4] proposed an alternative condition, called *recoverable linearizability*, which achieves locality but may sacrifice program order after a crash. It is a relaxed version of persistent atomicity, which requires the operation to be linearized or aborted before any subsequent linearization by the pending thread on that same object.

Izraelevitz et al. [15] considered a real-world failure model, in which processes are assumed to fail together, as part of a full-system crash. Under this model, persistent atomicity and recoverable linearizability are indistinguishable (and thus local). The term *durable linearizability* was used to refer to this merged consistency condition under the restricted failure model.

3. RECOVERABLE BASE OBJECTS

The correctness properties variants discussed in section 2.2 does not need to consider primitives, as any atomic operation satisfies any of the properties. Unlike it, RR-linearizability is not satisfied by primitives since the operation's response is lost in case of a failure. Therefore, there is a need to supply each primitive with an implementation as well as a recovery code. A system which supply recoverable primitives can be used to implement any object in a recoverable way - in case of a crash the process will simply recover the last atomic operation it executed before the crash. In case the operation is aborted, the process will reissue it. After that, it can continue and execute the remaining code safely. This observation focus our attention to implementation of primitives in a recoverable manner.

In the following section we present RR-linearizable implementations for well known primitives, as well as presenting impossibility result for others. We focus our attention on bounded wait-free implementations, that is, the number of steps a process takes when executing the recovery code in the absence of a failure is finite and bounded by a known constant (may be a function of n, the number of processes in the system), regardless of the other processes steps and the failures the process experienced so far. In addition, we would like the recovery code to use a finite number of variables. The fact that RR-linearizabilty allows us to swift the linearization point of an operation to after the crash is used to recover after a primitive failure.

Read. The process simply aborts the read upon recovering. Since read does not affect other processes, aborting the operation does not damage the linearizability of the program.

Write. Write instruction is "wrap" with code such that in case of a failure the extra data will be used for recovering. For an instruction writing value x to variable R by process p, we provide the following implementation. We use the convention of capital letters names for shared memory variables, and small letters for local ones. In the following code, R_p is a variable in the memory designated for process p.

Algorithm 1 Write

1: **procedure** WRITE OPERATION
2:     $res \leftarrow R$
3:     $R_p \leftarrow res$
4:     $R \leftarrow x$
5: **procedure** RECOVERY CODE
6:     **if** $R_p == R$ **then return** abort

For simplicity, we write the recovery code as a single instruction, although it needs to be written as several instructions, as it accesses two different locations in the shared memory. Since R_p is accessed by p only, the point where p reads R determines the recovery code outcome. Moreover, in case of a failure along the recovery code, the process can simply restart it, as it contains only reads.

The intuition for correctness is that if there was a write to R between the two reads of p (at line 2, and at the recovery code), then either this write is by p, and we can linearize it at the point where it took affect, or that there was a write by some other process, and we can linearize the write of p just before it. Hence, the real write "overwrite" p's write, and the rest of the processes can not distinguish between the two situations. Therefore, in case of a failure before line 4, p will simply abort upon recovering, and in case of a failure at line 4, p will execute the recovery code.

The above analysis ignores the ABA problem. It might be that p reads the same value from R, even though there was a write to R in between the two different reads. To overcome this problem, we can augment any value written with the writing process's id, and a sequential number (each process will have its own sequential number). This way, reading the same value guarantee that no write to R took place between the two different reads.

Compare-and-Swap. A Compare-and-Swap (CAS) object supports a single operation which atomically compares the value of the shared variable with its first parameter, and if they are equal, sets the value of the variable to its second parameter. At a high level, a process p first reads the CAS variable. If it observes a value different then *old*, then it return false. In such case, the operation can be linearized at the time of the read. Otherwise, p announce the process which is value was stored in the CAS object, that it reads its value, by writing to a designated memory, and only then it can try and apply the CAS operation to the object. This way, if p fails after a successful CAS operation, a different process that wants to change the value of the CAS, first needs to announce p his CAS was successful. Therefore, upon recovery, p can identify if its CAS took affect by read-

ing C, and looking for an info by different process that have seen p's value.

Algorithm 2 Compare-and-Swap

1: **procedure** CAS OPERATION
2: $< id, val > \leftarrow C$
3:    **if** $val \neq old$ **then**
4:       **return** false
5:    $R[id][i] \leftarrow val$
6:    $res \leftarrow C.cas(< id, val >, < i, new >)$
7: **procedure** RECOVERY CODE
8:    Read C, R[i][*]
9:    **if** $< id, val >$ appears in C, or val appears in R[i][*] **then**
10:       **return** true
11:    **else**
12:       **return** false

---

4. DISCUSSION

5. REFERENCES

[1] M. K. Aguilera and S. Frølund. Strict linearizability and the power of aborting. In *Tech. Rep. HPL-2003-241*, 2003.

[2] H. Attiya, R. Guerraoui, D. Hendler, and P. Kuznetsov. The complexity of obstruction-free implementations. *J. ACM*, 56(4):24:1–24:33, 2009.

[3] H. Attiya, D. Hendler, and P. Woelfel. Tight RMR lower bounds for mutual exclusion and other problems. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 217–226, 2008.

[4] R. Berryhill, W. M. Golab, and M. Tripunitara. Robust shared objects for non-volatile main memory. In *19th International Conference on Principles of Distributed Systems, OPODIS*, pages 20:1–20:17, 2015.

[5] H.-J. Boehm and D. R. Chakrabarti. Persistence programming models for non-volatile memory. In *Proceedings of the 2016 ACM SIGPLAN International Symposium on Memory Management*, ISMM, pages 55–67, 2016.

[6] J. Coburn, A. M. Caulfield, A. Akel, L. M. Grupp, R. K. Gupta, R. Jhala, and S. Swanson. NV-Heaps: Making persistent objects fast and safe with next-generation, non-volatile memories. In *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS, pages 105–118, 2011.

[7] M. Friedman, M. Herlihy, V. Marathe, and E. Petrank. Brief announcement: A persistent lock-free queue for non-volatile memory. In *Distributed Computing - 31st International Symposium (DISC)*, 2017.

[8] W. M. Golab and D. Hendler. Recoverable mutual exclusion in sub-logarithmic time. In *Proceedings of the ACM Symposium on Principles of Distributed Computing, PODC*, pages 211–220, 2017.

[9] W. M. Golab and A. Ramaraju. Recoverable mutual exclusion: [extended abstract]. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing, PODC*, pages 65–74, 2016.

[10] R. Guerraoui and R. R. Levy. Robust emulations of shared memory in a crash-recovery model. In *24th International Conference on Distributed Computing Systems (ICDCS)*, pages 400–407, 2004.

[11] M. Herlihy. Wait-free synchronization. *ACM Trans. Program. Lang. Syst.*, 13(1):124–149, 1991.

[12] M. Herlihy, V. Luchangco, and M. Moir. Obstruction-free synchronization: Double-ended queues as an example. In *23rd International Conference on Distributed Computing Systems (ICDCS))*, pages 522–529, 2003.

[13] M. Herlihy, V. Luchangco, M. Moir, and W. N. S. III. Software transactional memory for dynamic-sized data structures. In *Proceedings of the Twenty-Second ACM Symposium on Principles of Distributed Computing, PODC*, pages 92–101, 2003.

[14] M. Herlihy and J. M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, 1990.

[15] J. Izraelevitz, H. Mendes, and M. L. Scott. Linearizability of persistent memory objects under a full-system-crash failure model. In *Distributed Computing - 30th International Symposium, DISC*, pages 313–327, 2016.

[16] P. Jayanti and A. Joshi. Recoverable FCFS mutual exclusion with wait-free recovery. In *Distributed Computing - 31st International Symposium (DISC)*, 2017.

[17] A. Kolli, S. Pelley, A. Saidi, P. M. Chen, and T. F. Wenisch. High-performance transactions for persistent memories. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS, pages 399–411, 2016.

[18] L. Lamport. How to make a correct multiprocess program execute correctly on a multiprocessor. *IEEE Trans. Computers*, 46(7):779–782, 1997.

[19] V. J. Marathe, M. Seltzer, S. Byan, and T. Harris. Persistent memcached: Bringing legacy code to byte-addressable persistent memory. In *9th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage)*, 2017.

[20] M. M. Michael and M. L. Scott. Simple, fast, and practical non-blocking and blocking concurrent queue algorithms. In *Proceedings of the fifteenth annual ACM symposium on Principles of distributed computing*, pages 267–275, 1996.

[21] S. Venkataraman, N. Tolia, P. Ranganathan, and R. H. Campbell. Consistent and durable data structures for non-volatile byte-addressable memory. In *FAST*, volume 11, pages 61–75, 2011.

[22] H. Volos, A. J. Tack, and M. M. Swift. Mnemosyne: Lightweight persistent memory. In *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS, pages 91–104, 2011.