UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

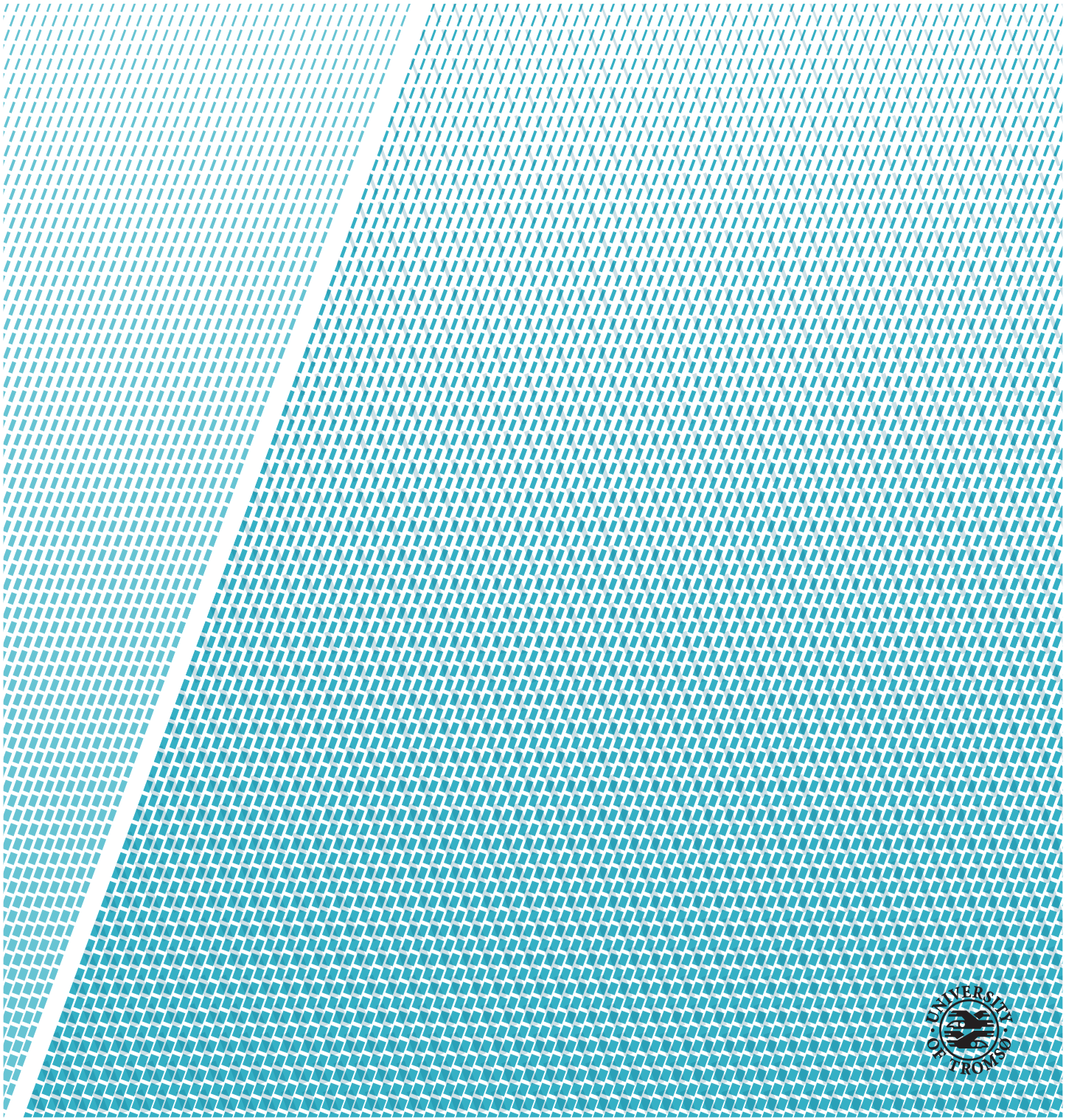
Faculty of Science and Technology
Department of Computer Science

Formal Verification of a Lock-free Split-order Hashmap

—

Åsmund Aqissiaq Arild Kløvstad

INF-2990 Bachelor's thesis in Informatics March 10, 2020



Contents

List of Figures

1 Introduction

Concurrent and distributed systems are extremely important in modern software development. Due to the difficulty of developing ever smaller and more powerful CPUs the trend in hardware design since about 2005 has been to increase the number of cores to allow for high levels of parallelization[1]. Additionally important areas of computing such as image processing and machine learning lend themselves well to such parallelization. [citation needed, Phuong?] At the same time software as a service and the massive scale industry giants like Amazon require a complex network of distributed systems to provide their functionality robustly and efficiently[2].

Hash tables are an important data structure for a variety of applications because they allow for data retrieval in constant time. Several lock-based hash tables for concurrent systems exist [citations], but the overhead of lock management and difficulty of resizing often make these impractical or inefficient[3]. A lock free alternative is proposed by Shalev and Shavit in [3]. This approach has proven to be useful[4] and scale better than lock-based approaches[5].

In both small- and large-scale computer systems it is important to ensure correctness. This is especially evident in critical infrastructure, but all scales and importance levels benefit from confidence in the correctness of their systems. [citation needed]

It is therefore troublesome that such systems are incredibly difficult to design, debug and reason about. The complexity of interactions between processes

and sheer number of possible edge cases makes it infeasible for a person to determine correctness.

Early solutions to this problem include Hoare[6], Floyd[7] and Pnueli's[8] temporal logics and Leslie Lamport's Temporal Logic of Actions[9] which seek to formalize the execution of programs in order to reason about them with logic. These formal methods proved useful, but laborious[10].

Building on the work in temporal logics, model checkers seek to minimize the human labor and ingenuity needed to prove correctness. This is done by specifying a model using some system of logic and then letting a model checker exhaustively survey the possible states of the system. This automates the process of proving correctness. One such model checker is the TLC model checker based on Lamport's TLA and incorporated in the TLA+ IDE.

1.1 Thesis

Shalev et al.'s split-ordered list design is a correct extensible hashmap for concurrent systems.

Furthermore it is possible to check this using a formal model checker, and the results of this will correspond to the properties proven by Shalev et al.

1.2 Method

In order to prove the correctness of the hashmap, its behavior will be implemented as a specification in TLA+[11] and the TLC model checker will be used to test the claimed invariants.

The specification will be developed in stages of increasing granularity, assuming the atomicity of operations to begin with and gradually loosening assumptions.

1.3 Scope?

1.4 Outline

Chapter 2 discusses the motivation for formal verification and model checking, followed by a description of Temporal Logic in Section 2.1 and the TLA+ language in Section 2.2. Finally Shalev et al.'s hashmap design is described in Section 2.3.

Chapter 3 describes the specification of the hashmap in TLA+ and the development of this specification.

Finally **Chapter 4** describes the results of model checking and discusses these in relation to our thesis in Section 4.1.

2

Background

Model Checking: Algorithmic Verification and Debugging[10] In the Turing Lecture by the winners of the 2007 Turing Award, Edmund Clarke, Allen Emerson and Joseph Sifakis they describe the develpment and use of model checkers as a verification method for computer systems. Previous efforts to prove correctness had been focused on formal proofs which have three key shortcomings:

1. they require human ingenuity,
2. they are difficult to work with in concurrent and distributed systems,
3. they scale poorly with system size and complexity.

Instead, they propose algorithmic model checkers.

With this method a Temporal Logic is used to specify the correct behavior of a system and the model checker verifies that this behavior is not violated by exploring the state space of the model. Importantly, such model checkers produce a counter example – an example of incorrect behavior – which makes debugging and correcting the system easier. Key properties of a temporal logic are *expressiveness* and *efficiency*.

Model checking also scales poorly with system complexity, so several techniques are introduced to deal with "state space explosion"

- symbolic checking of ordered binary decision diagrams
- isolation of independent events in concurrent systems
- bounded checking by solving SAT
- reduce state space by increasing level of abstraction
  - if counterexamples are found a lower abstraction level is needed, but "good" properties hold through abstraction mappings

How Amazon Web Services Uses Formal Methods[2] Amazon's AWS services are all underpinned by large and complex distributed systems. This is necessary for high availability, growth and cost-effective infrastructure. Traditionally these systems have been tested by savvy engineers who know what to test and look for. However, some errors are very rare and will very likely slip through such testing. To catch these errors they employ model checking (with TLA+).

The PlusCal or TLA+ specifications work as a tool to bridge the gap between design and implementation. Designs are expressive, but inprecise while the implementation is precise, but hides overall structure. Through a choice of abstraction level, specifications can bridge this gap and provide both. An expressive specification also provides useful documentation of the system.

The key benefits of model checkers at Amazon are:

- a precisely specified design helps make changes and optimizations safely. This usage improves system understanding.
- they are faster than formal proofs
- a correct design and the understanding the specs provide promote better, more correct code.

2.1 Temporal Logics

The Temporal Logic of Programs [8] In The Temporal Logic of Programs[8], Amir Pnueli proposes a unified approach to the verification of both sequential and concurrent programs. His work seeks to unify approaches to to both, while also presenting a system that emulates the design intuition of programmers. The key concepts in this work are *invariance* – which covers par-

tial correctness, clean behavior, mutual exlusion and deadlock freedom – and *eventuality* – which generalizes these notions to cyclic programs and provides a special case of total correctness.

A dynamic discrete system is generalized as a three-tuple $\langle S, R, s_0 \rangle$ where S is the set of possible states, R a transition relation, and s_0 the initial state of the system. In order to make later constructions easier we further specify

$$s = \langle \pi, u \rangle$$

where π is the control component specifying the location in the program and u is the data component describing the state of any variables and data structures, and

$$R(\pi, u) = N(\pi, u) \wedge T(\pi, u)$$

where N describes the control flow and T the change in data such that a step in the execution may be described by

$$R(\langle \pi, u \rangle, \langle \pi', u' \rangle) \iff \pi' = N(\pi, u) \wedge u' = T(\pi, u)$$

To reason about concurrent programs we let states have multiple control components $s = \langle \pi_1, \pi_2, ..., \pi_n, u \rangle$ and randomly choose one control component to update in each step. Finally we let X be the set of all reachable states for the system. A predicate $p(s)$ is **invariant** if $p(s)$ is true $\forall s \in X$.

We can now start to define useful properties of the systems described in this way.

> **Partial correctness** is the claim that given the correct input, a program produces the correct output. We let $\phi(x)$ be the statement "reaching the end state $\rightarrow$ (correct input $\rightarrow$ correct output)". Partial correctness is equivalent to saying ϕ is an invariant.
>
> **Clean execution** means the program does not behave illegally, i.e it does not access illegal memory locations or divide by zero. We may define these restrictions as a predicate to make clean execution equivalent to this predicate being invariant.
>
> **Mutual exclusion**. Given a critical section C, mutual exclusion of the processes π_1 and π_2 is described by the invariance of the predicate $\neg(\pi_1 \in C \wedge \pi_2 \in C)$.

In addition to these properties we wish to reason about *temporal* implications. We let time be described by a $t \in \mathbb{N}$ and $H(p, t)$ denote the value of the predicate p at time t. We then introduce the temporal operator $p \rightsquigarrow q$ to mean

p eventually leads to q, or formally:

$$p \rightsquigarrow q : \forall t_1 \exists t_2 \; s.t \; t_1 \leq t_2, H(p, t_1) \rightarrow H(q, t_2)$$

For all times t_1 there is a later time t_2 such that if p holds at t_1, q will hold at t_2. Armed with eventuality we can define temporally useful properties of systems.

Total correctness is stronger than partial correctness because it also requires that the program reaches an end state. We can express total correctness as $\langle \pi = l_0, u = \phi \rangle \rightsquigarrow \langle \pi = l_m, u = \psi \rangle$ where ϕ denotes correct input, and ψ denotes correct output and l_0, l_m are the start and end labeles of the system, respectively.

Accessibility is the guarantee that some segment S of a program can be reached. It can be expressed by $\pi = l_0 \rightsquigarrow \pi \in S$

Responsiveness. It is often desirable that some request r will be met by a response s. We call this responsiveness and describe it by $r \rightsquigarrow s$.

With these definitions under our belt, Pnueli defines the necessary axioms and inference rules to reason about the correctness of programs.

$$[\forall s, s' p(s) \wedge R(s, s') \rightarrow q(s')] \Rightarrow p \rightsquigarrow q \tag{A1}$$

$$(p \rightarrow q) \Rightarrow p \rightsquigarrow q \tag{A2}$$

Figure 2.1: Pnueli's axioms

These axioms define two ways to establish eventuality. A2 says that any logical implication is also an eventuality. A1 is a little more involved, but states that if for all consecutive states p being true in the first implies q being true in the second, then p eventually leads to q.

Using these axioms and inference rules as well as first-order logic, Pnueli goes on to formalize invariance and eventuality for sequential programs and concurrent programs.

Invariance of the predicate $q(\pi, u)$ is described by the conjunction $\bigwedge_i \pi = l_i \rightarrow q(l_i, u)$, asserting that q is true at all points of execution. This method is called an *attachment* of the predicate to the program.

$$p \rightsquigarrow q, \forall s, s' r(s) \wedge R(s, s') \rightarrow r(s') \Rightarrow (p \wedge r) \rightsquigarrow (q \wedge r) \quad \text{(R1)}$$

$$p \rightsquigarrow q, q \rightsquigarrow r \Rightarrow p \rightsquigarrow r \quad \text{(R2)}$$

$$p_1 \rightsquigarrow q, p_2 \rightsquigarrow q \Rightarrow (p_1 \vee p_2) \rightsquigarrow q \quad \text{(R3)}$$

$$p \rightsquigarrow q \Rightarrow (\exists u p) \rightsquigarrow q \quad \text{(R4)}$$

Figure 2.2: Pnueli's inference rules

In a concurrent program we generalize q to hold when any π_i is updated by N and construct either a full attachment

$$\bigwedge_{i_1, i_2, \ldots, i_n} (\pi_1 = i_1 \wedge \pi_2 = i_2 \wedge \ldots \wedge \pi_n = i_n) \rightarrow q(\pi_1, .., \pi_n, u)$$

shown here for n concurrent execution threads, or the partial attachment

$$\bigwedge_i \pi_1 = i \rightarrow q(\pi_1 = i, \pi_2, u) \wedge \bigwedge_j \pi_2 = j \rightarrow q(\pi_1, \pi_2 = j, u)$$

shown here for two threads π_1 and π_2.

Eventuality is formulated as the temporal implication $\pi = l_1 \wedge p(u) \rightsquigarrow \pi = l_2 \wedge q(u)$. We can then describe the path between l_1 and l_2 by a finite sequence of steps and apply A1 to each step.

Finally, Pnueli introduces two new "tense operators" **F**uture and **G**lobal on predicates such that at some time n

$$F(p) = \exists t \geq n \; s.t \; H(t, p)$$

$$G(p) = \forall t \geq n \; H(t, p)$$

This lets us describe useful properties such as

$p \rightarrow F(q)$ – if p is true now, then at some point in the future q will be true.

$G(p \rightarrow F(q))$ – whenever p is true it will eventually be followed by a state in which q is true. (this is equivalent to $p \rightsquigarrow q$)

2.2 TLA+

2.3 Split-Ordered List Hashmap

Maybe a description of hashmaps in general here to set up the later use of table, list, key etc. or is that assumed knowledge?

Shalev et al.[3] present the first lock-free extensible hash table implemented using only loads, stores and atomic Compare and Swap (CAS).

Hashmaps are a key building block in many important systems [citation needed], but are difficult to implement concurrently. In particular, the resizing (extending) of the table is difficult to do atomically because at the very least a node must be moved from one list to another. In order to avoid conflicts and loss of data in this process some overhead is required which impacts performance.

The key insight of Shalev et al. is to flip the process upside down. Instead of moving nodes between buckets, they suggest moving the buckets among a statically ordered list of nodes. This requires an ordering of the list in which a bucket can always be split into two new buckets while their contents remain correct. A node should always reside in the bucket corresponding to its key $\mod 2^i$ where 2^i is the current size of the table.

Split-Ordered Lists are introduced to solve this problem. By sorting the keys according to their reversed binary representation Shalev et al. obtain a list which can be always be split into buckets $\mod 2^i$. This is because such an ordering corresponds to difference in the keys' ith least significant bit, which is equivalent to having a different remainder $\mod 2^i$.

In order to deal with the problems caused by removing nodes pointed to by hash table entries dummy nodes with the bucket value are introduced. These nodes signify the start of a bucket and are recursively initialized when an item is inserted into an unitialized bucket. To distinguish dummy nodes from regular nodes in the list, regular node keys have their most significant bit set to 1 before being reversed.

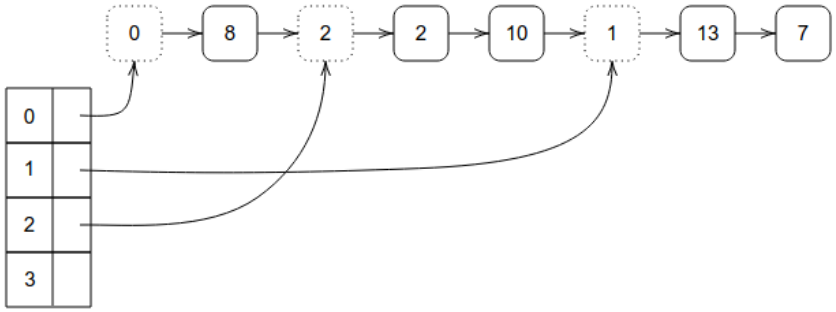
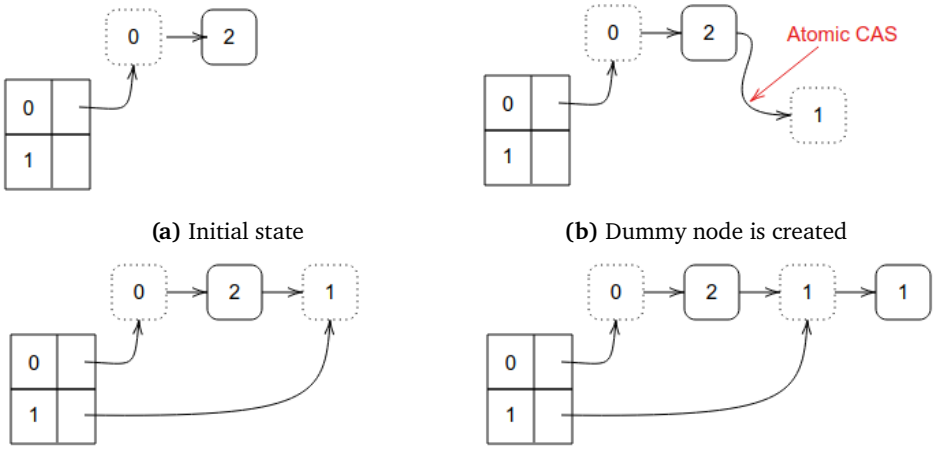


Figure 2.3: The layout of the split-ordered list



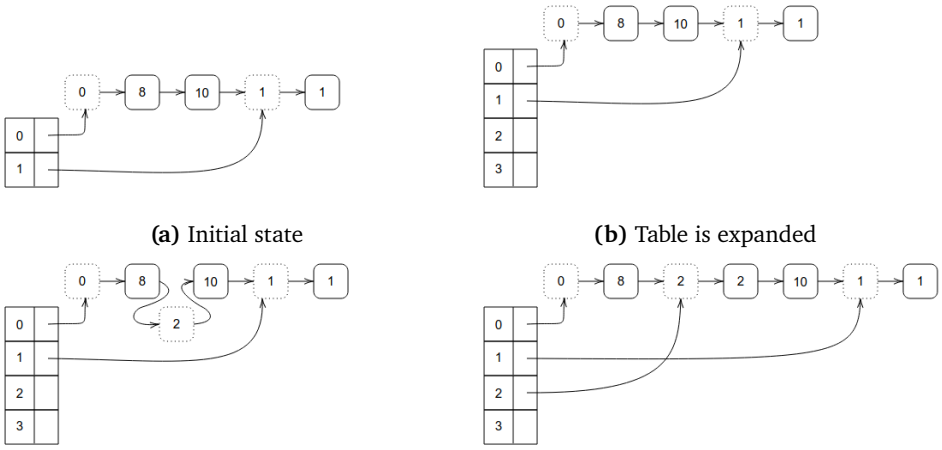
(a) Initial state

(b) Dummy node is created

(c) Dummy node is inserted in bucket

(d) New node is inserted

Figure 2.4: Insertion without bucket splitting



(a) Initial state

(b) Table is expanded

(c) Dummy node for new bucket inserted

(d) Table entry points to bucket node

Figure 2.5: Expansion and bucket splitting

3

TLA+ Specification

4

Results

4.1 Discussion

Bibliography

[1] Andrew S. Tanenbaum and Herbert Bos. *Modern Operating Systems*. 4th. USA: Prentice Hall Press, 2014. ISBN: 013359162X.

[2] Chris Newcombe et al. "How Amazon Web Services Uses Formal Methods." In: *Commun. ACM* 58.4 (Mar. 2015), pp. 66–73. ISSN: 0001-0782. DOI: 10.1145/2699417. URL: https://doi.org/10.1145/2699417.

[3] Ori Shalev and Nir Shavit. "Split-Ordered Lists: Lock-Free Extensible Hash Tables." In: *J. ACM* 53.3 (2006), pp. 379–405. ISSN: 0004-5411. DOI: 10.1145/1147954.1147958. URL: https://doi.org/10.1145/1147954.1147958.

[4] Daniel Cederman et al. *Lock-free Concurrent Data Structures*. 2013. arXiv: 1302.2757 [cs.DC].

[5] Rodrigo Medeiros Duarte et al. "Concurrent Hash Tables for Haskell." In: *Programming Languages*. Ed. by Fernando Castor and Yu David Liu. Cham: Springer International Publishing, 2016, pp. 110–124. ISBN: 978-3-319-45279-1.

[6] C. A. R. Hoare. "An Axiomatic Basis for Computer Programming." In: *Commun. ACM* 12.10 (Oct. 1969), pp. 576–580. ISSN: 0001-0782. DOI: 10.1145/363235.363259. URL: https://doi.org/10.1145/363235.363259.

[7] Robert W. Floyd. "Assigning Meanings to Programs." In: *Proceedings of Symposium on Applied Mathematics* 19 (1967), pp. 19–32. URL: http://laser.cs.umass.edu/courses/cs521-621.Spr06/papers/Floyd.pdf.

[8] Amir Pnueli. "The Temporal Logic of Programs." In: *Proceedings of the 18th Annual Symposium on Foundations of Computer Science*. SFCS '77. USA: IEEE Computer Society, 1977, pp. 46–57. DOI: 10.1109/SFCS.1977.32. URL: https://doi.org/10.1109/SFCS.1977.32.

[9] L Lamport. "Proving the Correctness of Multiprocess Programs." In: *IEEE Trans. Softw. Eng.* 3.2 (1977), pp. 125–143. ISSN: 0098-5589. DOI: 10.1109/TSE.1977.229904. URL: https://doi.org/10.1109/TSE.1977.229904.

[10] Edmund M. Clarke, E. Allen Emerson, and Joseph Sifakis. "Model Checking: Algorithmic Verification and Debugging." In: *Commun. ACM* 52.11 (Nov. 2009), pp. 74–84. ISSN: 0001-0782. DOI: 10.1145/1592761.1592781. URL: https://doi.org/10.1145/1592761.1592781.

[11] Leslie Lamport et al. “Specifying and Verifying Systems with TLA+.” In: *Proceedings of the 10th Workshop on ACM SIGOPS European Workshop*. EW 10. New York, NY, USA: ACM, 2002, pp. 45–48. DOI: 10.1145/1133373.1133382. URL: http://doi.acm.org/10.1145/1133373.1133382.