

Zoology of lockfree concurrent data structures

Clément Allain

INRIA Paris

June 18, 2024

Zoology

Specimen ① : Michael-Scott queue

Specimen ② : KCAS

Zoology: what is it?

HEAPLANG (modified)

- + ADTs
- + DLS
- + exceptions
- + algebraic effects
- + relaxed memory

(planned before OSIRIS,
in case you were wondering)

Zoo

IRIS

CoQ

Zoology: what is it?

```
MetaCoq Run (zoo_variant "list" [  
    "Nil" ;  
    "Cons"  
]).
```

```
Definition map : val :=  
  rec: "map" "fn" "xs" :=  
    match: "xs" with  
      | Nil =>  
        §Nil  
      | Cons "x" "xs" =>  
        let: "y" := "fn" "x" in  
        'Cons{ "y", "map" "fn" "xs" }  
    end.
```

IRIS

CoQ

Zoology: what is it?

HEAPLANG (modified)

- + ADTs
- + DLS
- + exceptions
- + algebraic effects
- + relaxed memory

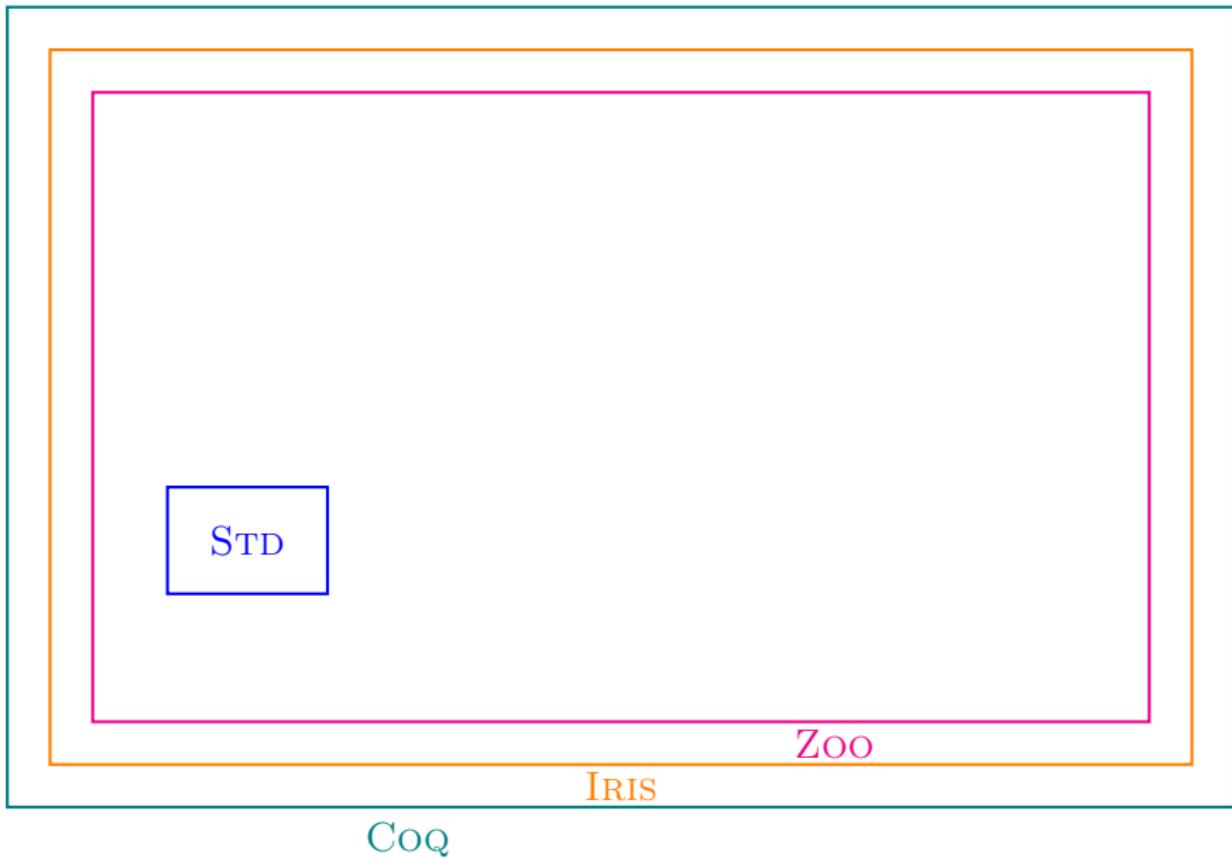
(planned before OSIRIS,
in case you were wondering)

Zoo

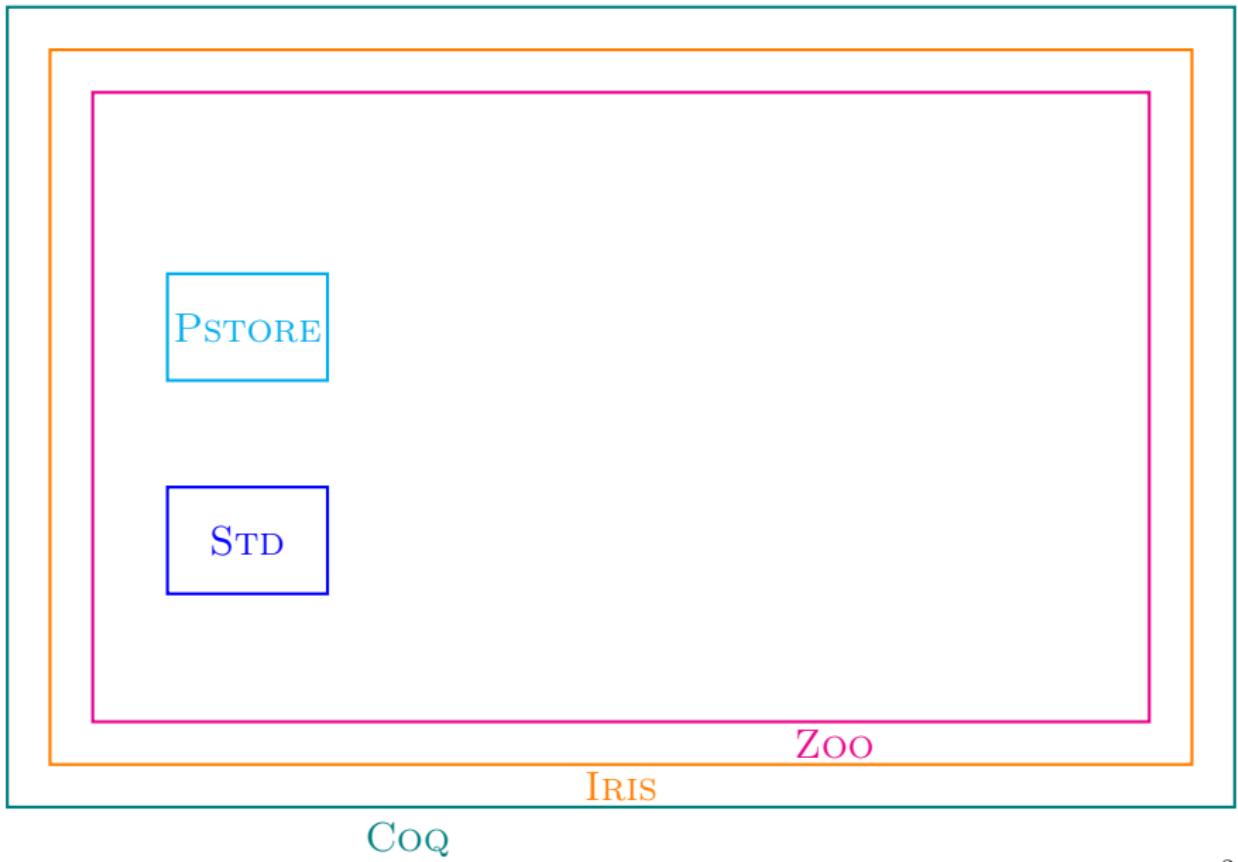
IRIS

CoQ

Zoology: what is it?



Zoology: what is it?



Zoology: what is it?



Clément
Allain



Basile
Clément



Alexandre
Moine



Gabriel
Scherer

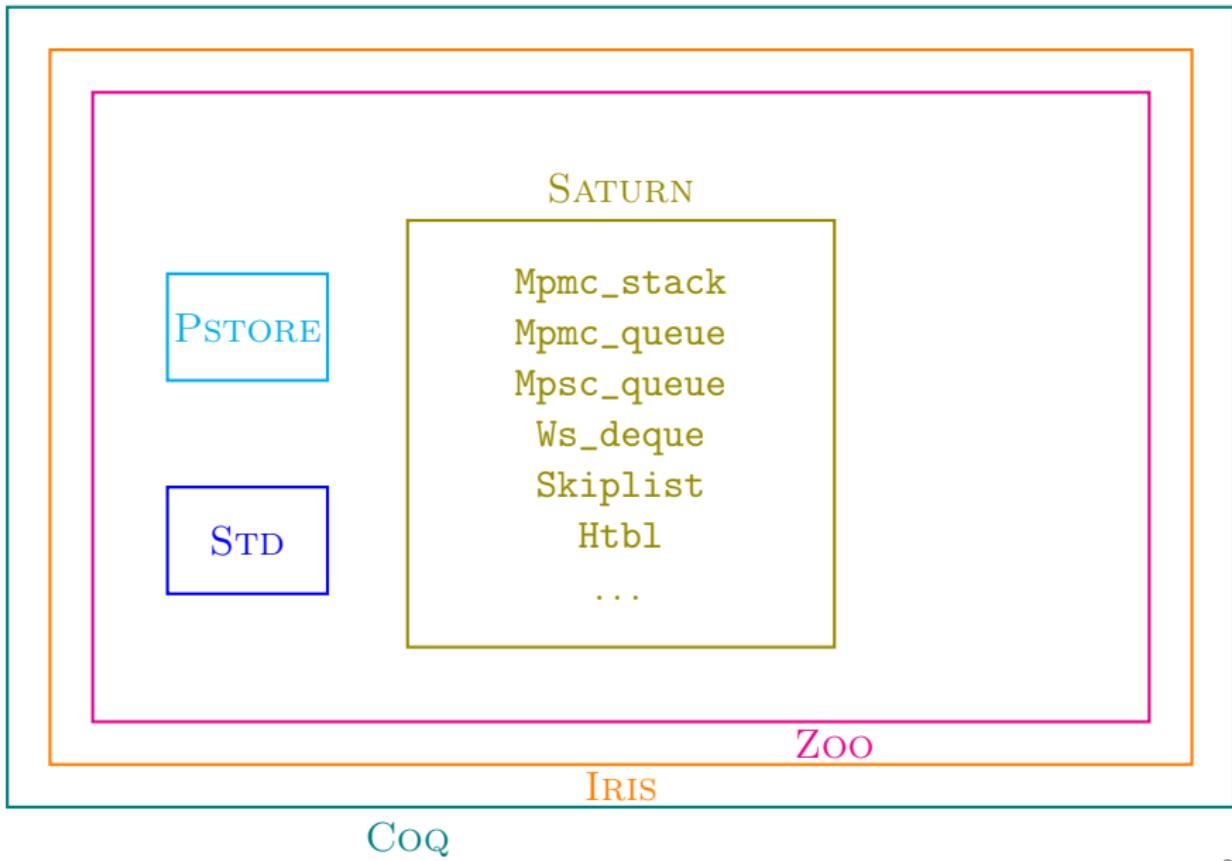
The PSTORE team!

Zoo

IRIS

CoQ

Zoology: what is it?



Zoology: what is it?



Vesa Karvonen



Carine Morel

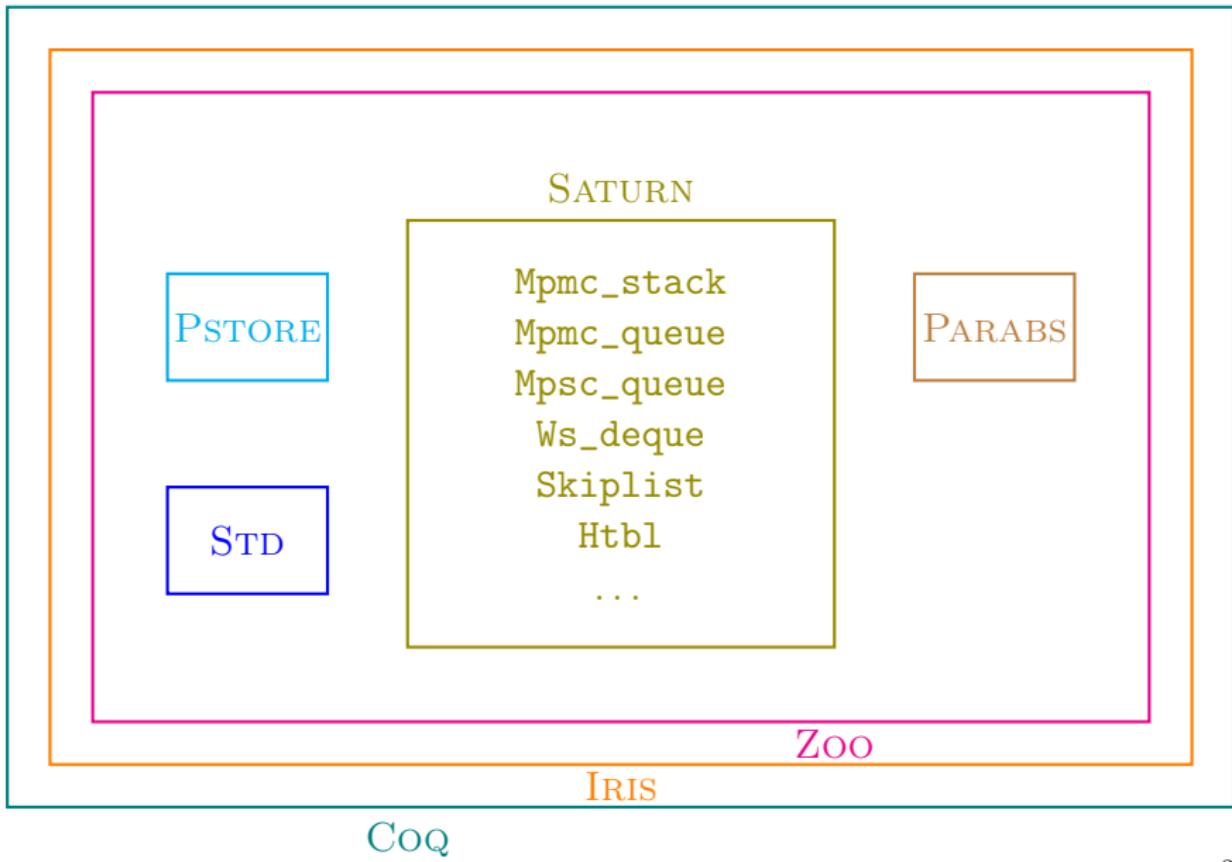
The SATURN team!

Zoo

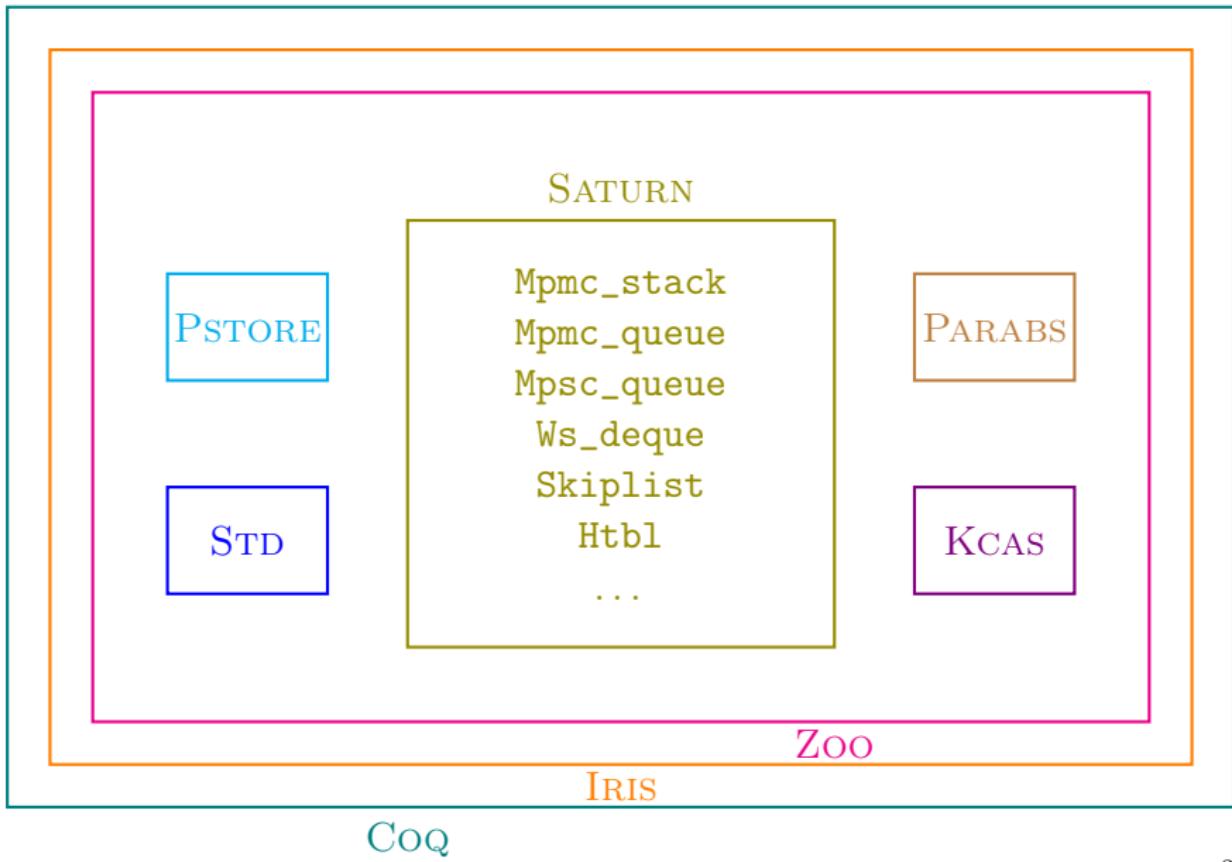
IRIS

CoQ

Zoology: what is it?



Zoology: what is it?



Zoology: what is it?



Vesa Karvonen

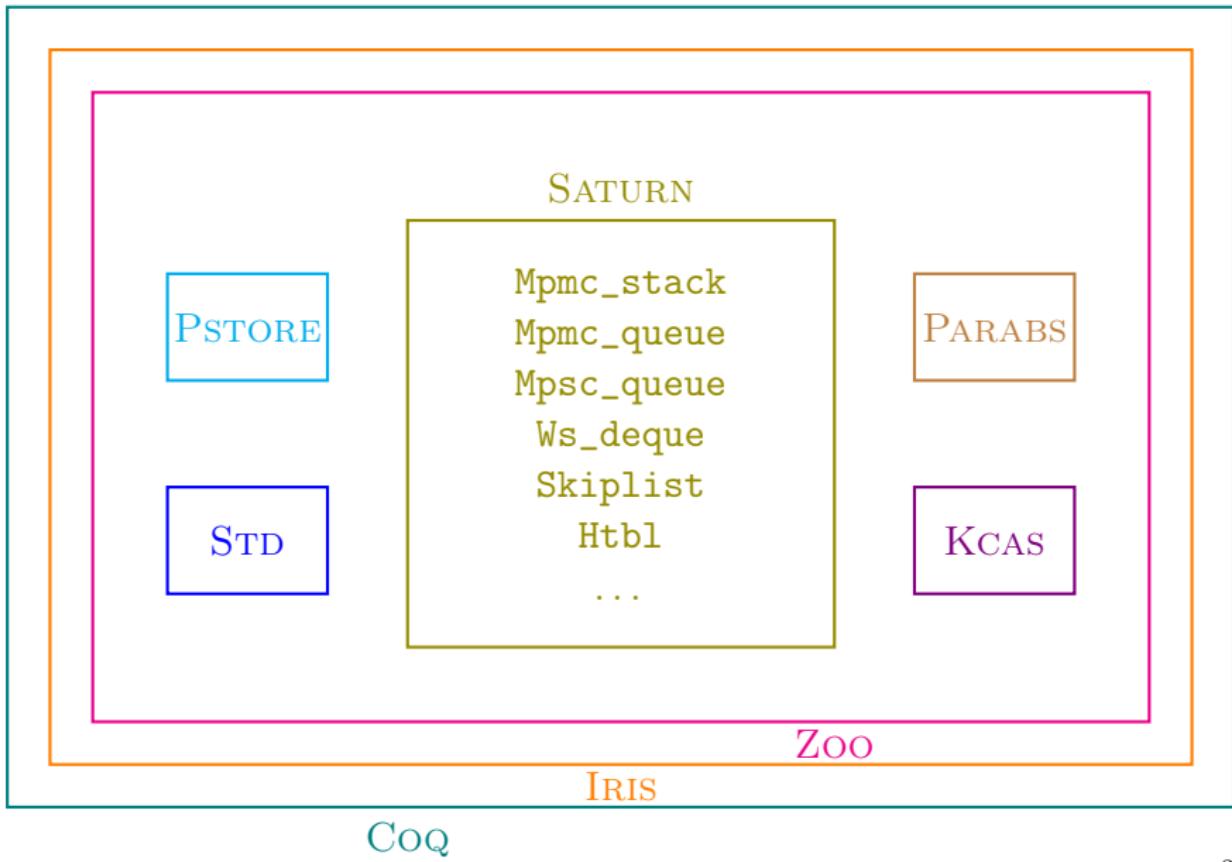
The main author of KCAS!

Zoo

IRIS

CoQ

Zoology: what is it?



Zoology: why is it fun?

Lockfree algorithms typically exhibit complex behaviors:

- ▶ physical state \neq logical state,
- ▶ external linearization points,
- ▶ future-dependent linearization points.

IRIS is a good match for verifying them thanks to advanced mechanisms:

- ▶ invariants to enforce protocols,
- ▶ atomic updates to materialize linearization points,
- ▶ prophecy variables to reason about the future.

Zoology

Specimen ① : Michael-Scott queue

Specimen ② : KCAS

Original algorithm (1996)

Simple, Fast, and Practical Non-Blocking and Blocking Concurrent Queue Algorithms*

Maged M. Michael Michael L. Scott

Department of Computer Science

University of Rochester

Rochester, NY 14627-0226

{michael,scott}@cs.rochester.edu

Abstract

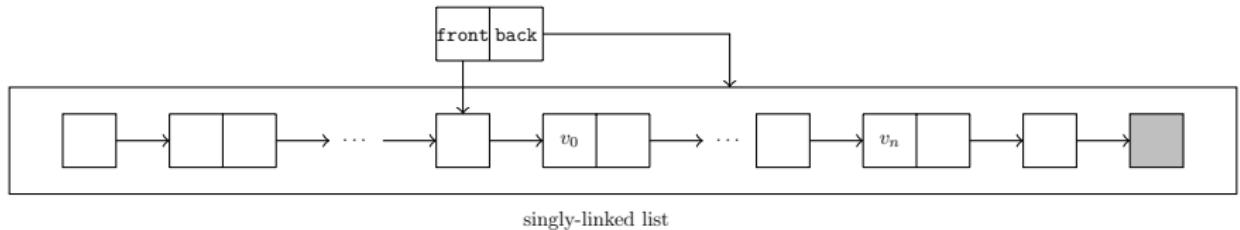
Drawing ideas from previous authors, we present a new non-blocking concurrent queue algorithm and a new two-lock queue algorithm in which one enqueue and one dequeue can proceed concurrently. Both algorithms are simple, fast, and practical; we were surprised not to find them in the literature. Experiments on a 12-node SGI Challenge multiprocessor indicate that the new non-blocking queue consistently outperforms the best known alternatives; it is the clear algorithm of choice for machines that provide a universal atomic primitive (e.g. `compare_and_swap` or `load_linked/store_conditional`). The two-lock concurrent queue outperforms a single lock when several processes are competing simultaneously for access; it appears to be the algorithm of choice for busy queues on machines with non-universal atomic primitives (e.g. `test_and_set`). Since much of the motivation for non-blocking algorithms is rooted in their immunity to large, unpredictable delays in process execution, we report experimental results both for systems with dedicated processors and for systems with shared memory. The results show that the new non-blocking queue is significantly faster than the best known alternative, especially when the queue is busy.

1 Introduction

Concurrent FIFO queues are widely used in parallel applications and operating systems. To ensure correctness, concurrent access to shared queues has to be synchronized. Generally, algorithms for concurrent data structures, including FIFO queues, fall into two categories: *blocking* and *non-blocking*. Blocking algorithms allow a slow or delayed process to prevent faster processes from completing operations on the shared data structure indefinitely. Non-blocking algorithms guarantee that if there are one or more active processes trying to perform operations on a shared data structure, some operation will complete within a finite number of time steps. On asynchronous (especially multiprogrammed) multiprocessor systems, blocking algorithms suffer significant performance degradation when a process is halted or delayed at an inopportune moment. Possible sources of delay include processor scheduling preemption, page faults, and cache misses. Non-blocking algorithms are more robust in the face of these events.

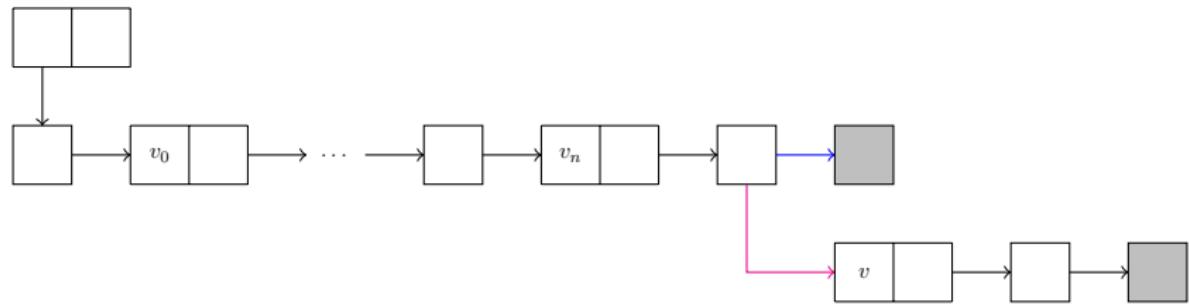
Many researchers have proposed lock-free algorithms for concurrent FIFO queues. Hwang and Briggs [7],

Basic implementation



- ▶ The queue contains $[v_0, \dots, v_n]$.
- ▶ `front` points to the first component of the active part of the list.
- ▶ `back` points somewhere in the list, usually closer to the end.

push linearization

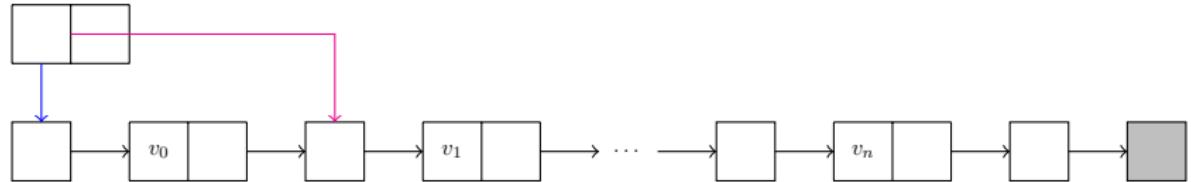


queue-model $t [v_0, \dots, v_n]$



queue-model $t [v_0, \dots, v_n, v]$

pop linearization



queue-model t $[v_0, \dots, v_n]$
↓
queue-model t $[v_1, \dots, v_n]$

Basic OCAML 5 implementation

```
type 'a node =
| Nil
| Next of 'a * 'a node Atomic.t

type 'a t = {
  front: 'a node Atomic.t Atomic.t ;
  back: 'a node Atomic.t Atomic.t ;
}
```

Specification

$$\frac{\frac{\frac{\{ \text{queue-inv } t \iota \}}{\langle \forall vs. \text{queue-model } t vs \rangle}}{\text{queue_push } t v, \uparrow \iota}}{\langle \text{queue-model } t (vs \uplus [v]) \rangle}}{\{ () . \text{True} \}}$$

$$\frac{\frac{\frac{\{ \text{queue-inv } t \iota \}}{\langle \forall vs. \text{queue-model } t vs \rangle}}{\text{queue_pop } t, \uparrow \iota}}{\langle \text{queue-model } t (\text{tail } vs) \rangle}}{\{ \text{head } vs. \text{True} \}}$$

Verification by Vindum & Birkedal (2021)



Contextual Refinement of the Michael-Scott Queue (Proof Pearl)

Simon Friis Vindum
vindum@cs.au.dk
Aarhus University
Aarhus, Denmark

Lars Birkedal
birkedal@cs.au.dk
Aarhus University
Aarhus, Denmark

Abstract

The Michael-Scott queue (MS-queue) is a concurrent non-blocking queue. In an earlier pen-and-paper proof it was shown that a simplified variant of the MS-queue contextually refines a coarse-grained queue. Here we use the Iris and ReLoC logics to show, for the first time, that the original MS-queue contextually refines a coarse-grained queue. We make crucial use of the recently introduced prophecy variables of Iris and ReLoC. Our proof uses a fairly simple invariant that relies on encoding which nodes in the MS-queue can reach other nodes. To further simplify the proof, we extend separation logic with a generally applicable *persistent points-to predicate* for representing immutable pointers. This relies on a generalization of the well-known algebra of fractional permissions into one of *discardable* fractional permissions. We define the persistent points-to predicate entirely inside the base logic of Iris (thus getting soundness “for free”).

We use the same approach to prove refinement for a variant of the MS-queue resembling the one used in the `java.util.concurrent` library.

We have mechanized our proofs in Coq using the formalizations of ReLoC and Iris in Coq.

CCS Concepts: • Theory of computation → Separation logic; Concurrent algorithms.

Keywords: separation logic, concurrent vs. Iris, Coq

1 Introduction

The Michael-Scott queue (MS-queue) is a fast and practical fine-grained concurrent queue [14]. We prove that the MS-queue is a *contextual refinement* of a coarse-grained concurrent queue. The coarse-grained queue, shown in Figure 1, is implemented as a reference to a functional list and uses a lock to sequentialize concurrent accesses to the queue. We thus prove that in any program we may replace uses of the coarse-grained, but obviously correct, concurrent queue with the faster, but more intricate, MS-queue, without changing the observable behaviour of the program. We recall that, formally, an expression e contextually refines another expression e' , denoted $\Delta; \Gamma \vdash e \lesssim_{ctx} e' : \tau$, if for all contexts K , of ground type, whenever $K[e]$ terminates with a value there exists an execution of $K[e']$ that terminates with the same value. One should think of e as the *implementation* (in our case the MS-queue), e' as the *specification* (in our case the coarse-grained queue), and K as a *client* of a queue implementation.

Note that the contextual refinement implies that the internal states of the two queues are *encapsulated* and hidden from clients who could otherwise tell the difference between the two implementations. Contextual refinement is also related to *linearizability*, a popular correctness criterion considered for concurrent data structures. Linearizability has mostly been considered for first-order programming lan-



Contextual Refinement of the Michael-Scott Queue (Proof Pearl)

A key insight of our approach is how the invariants that the MS-queue maintains can be expressed in terms of which nodes are *reachable* from other nodes. Reachability is expressed with an inductive predicate:

$$\ell_n \rightsquigarrow \ell_m \triangleq \exists \ell_{n \rightarrow}, v. \ell_n \hookrightarrow_i^\square \text{some}(v, \ell_{n \rightarrow}) * \\ (\ell_n = \ell_m \vee \exists \ell_p. \ell_{n \rightarrow} \hookrightarrow_i^\square \ell_p * \ell_p \rightsquigarrow \ell_m)$$

separation logic with a generally applicable *persistent points-to predicate* for representing immutable pointers. This relies on a generalization of the well-known algebra of fractional permissions into one of *discardable* fractional permissions. We define the persistent points-to predicate entirely inside the base logic of Iris (thus getting soundness “for free”).

We use the same approach to prove refinement for a variant of the MS-queue resembling the one used in the `java.util.concurrent` library.

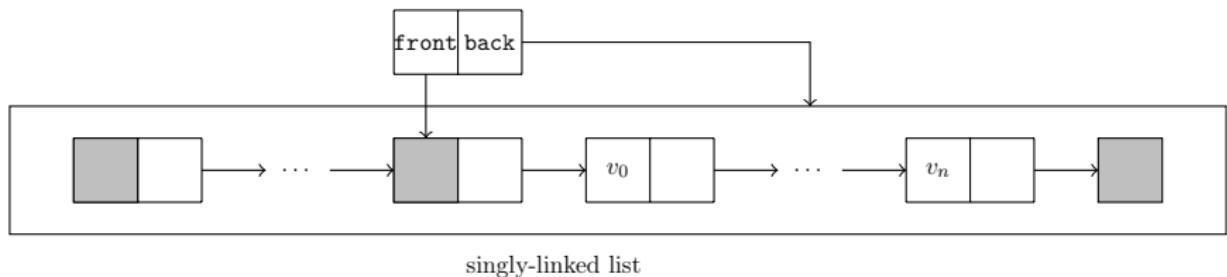
We have mechanized our proofs in Coq using the formalizations of ReLoC and Iris in Coq.

CCS Concepts: • Theory of computation → Separation logic; Concurrent algorithms.

formally, an expression e contextually refines another expression e' , denoted $\Delta; \Gamma \vdash e \lesssim_{ctx} e' : \tau$, if for all contexts K , of ground type, whenever $K[e]$ terminates with a value there exists an execution of $K[e']$ that terminates with the same value. One should think of e as the *implementation* (in our case the MS-queue), e' as the *specification* (in our case the coarse-grained queue), and K as a *client* of a queue implementation.

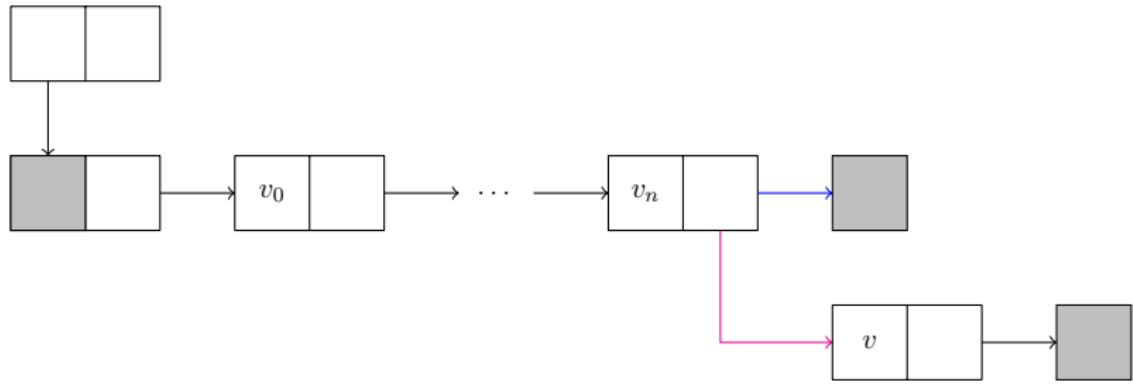
Note that the contextual refinement implies that the internal states of the two queues are *encapsulated* and hidden from clients who could otherwise tell the difference between the two implementations. Contextual refinement is also related to *linearizability*, a popular correctness criterion considered for concurrent data structures. Linearizability has mostly been considered for first-order programming lan-

Optimized implementation



- ▶ The queue contains $[v_0, \dots, v_n]$.
- ▶ **front** points to the first component of the active part of the list.
- ▶ **back** points somewhere in the list, usually closer to the end.

push linearization

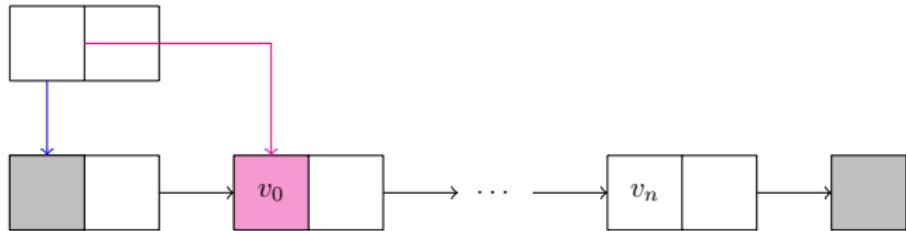


queue-model $t [v_0, \dots, v_n]$



queue-model $t [v_0, \dots, v_n, v]$

pop linearization



queue-model t $[v_0, \dots, v_n]$
 \Downarrow
queue-model t $[v_1, \dots, v_n]$

Optimized OCAML 5 implementation — first try

```
type ('a, _) node =
| Nil : ('a, [> `Nil]) node
| Next : {
    mutable next: ('a, [`Nil | `Next]) node ;
    mutable value: 'a ;
} ->
('a, [> `Next]) node

external node_as_atomic :
('a, [`Next]) node ->
('a, [`Nil | `Next]) node Atomic.t
= "%identity"

let push t v =
...
if Atomic.compare_and_set (node_as_atomic back) Nil node
then ...
else ...
```

A new hope

The screenshot shows a GitHub repository interface for the 'ocaml / RFCs' repository. The main navigation bar includes links for Code, Issues, Pull requests (21), Actions, Security, and Insights. The 'Pull requests' link is highlighted with an orange underline. Below the navigation, the title 'Atomic record fields #39' is displayed. A green button labeled 'Open' indicates the status of the pull request. The pull request details show it merging 2 commits from the user 'gasche' into the branch 'ocaml:master' from the 'gasche:atomic-record' repository. A small icon of a person with a checkmark is next to the merge message. Below the pull request summary, there are tabs for Conversation (9), Commits (2), Checks (0), and Files changed (1). A comment from the user 'gasche' is visible, dated 2 weeks ago and edited. The comment text reads: 'A proposed design for atomic record fields.'

(no modification of the OCAML 5 memory model)

Optimized OCAML 5 implementation — second try

```
type ('a, _) node =
| Nil : ('a, [> `Nil]) node
| Next of {
    mutable next: ('a, [`Nil | `Next]) node [@atomic] ;
    mutable value: 'a ;
} ->
('a, [> `Next]) node

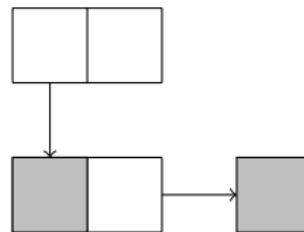
let push t v =
...
if Atomic.Loc.compare_and_set
[%atomic.field back.next]
Nil node
then ...
else ...
```

What about `is_empty`?

```
let is_empty t =  
  let Next front = t.front in  
  front.next == Nil
```

$$\frac{\frac{\frac{\frac{\{ \text{queue-inv } t \iota \} }{ \langle \forall vs. \text{queue-model } t vs \rangle } }{ \text{queue_is_empty } t, \uparrow \iota } }{ \langle \text{queue-model } t vs \rangle } }{ \left\{ vs \stackrel{?}{=} [] . \text{True} \right\} }$$

is_empty (potentially external) linearization



$$[v_0, \dots, v_n] = []$$

$$[v_0, \dots, v_n] \neq []$$



IRIS invariant

```
Definition queue_inv l γ ℓ : iProp Σ :=  
  ∃ hist past front nodes back vs waiters,  
  
  ▷ hist = past ++ front :: nodes ▷ *  
  ▷ back ∈ hist ▷ *  
  
  l.[front] ↪ #front *  
  l.[back] ↪ #back *  
  
queue_front_auth γ (length past) *  
  
queue_history_auth γ hist *  
xchain_model hist () *  
  
queue_model2 γ vs *  
([* list] node; v ∈ nodes; vs, node.[xchain_data] ↪ v) *  
  
queue_waiters_auth γ waiters *  
([* map] waiter ↪ i ∈ waiters, queue_waiter γ ℓ past waiter i).
```

What more?

- ▶ Multi-producer *single-consumer* queue.
- ▶ Multi-producer multi-consumer *bounded* queue.
- ▶ More advanced queues derived from Michael-Scott?
- ▶ Space complexity by Alexandre Moine.

Zoology

Specimen ① : Michael-Scott queue

Specimen ② : KCAS

TODO

Thank you for your attention!