Correct and Efficient Work-Stealing for Weak Memory Models

Nhat Minh Lê  Antoniu Pop  Albert Cohen  Francesco Zappa Nardelli
INRIA and ENS Paris

Abstract
Chase and Lev’s concurrent deque is a key data-structure in shared-memory parallel programming and plays an essential role in work-stealing schedulers. We provide the first correctness proof of an optimized implementation of Chase and Lev’s deque on top of the Power and ARMv7 architectures: these provide very relaxed memory models, which we exploit to improve performance but considerably complicate the formal reasoning. We also study an optimized x86 and a portable C11 implementation, conducting systematic experiments to evaluate the impact of memory barrier optimizations. Our results demonstrate the benefits of hand tuning the deque code when running on top of relaxed memory models.

Categories and Subject Descriptors D.1.3 [Programming Techniques]: Concurrent Programming; E.1 [Data Structures]: Lists, stacks, and queues

Keywords: lock-free algorithm, work-stealing, relaxed memory model, proof

1. Introduction
Multicore POWER and ARM architectures are standard targets for server, consumer electronics, and embedded control applications. The difficulties of parallel programming are exacerbated by the relaxed memory model implemented by these architectures, which allow the processors to perform a wide range of optimizations, including thread-local reordering and non-atomic store propagation.

The safety-critical nature of many embedded applications call for solid foundations for parallel programming. This paper shows that a high degree of confidence can be achieved for highly optimized, real-world, concurrent algorithms, running on top of weak memory models. A good test-case is provided by the runtime scheduler of a task library. We thus focus on the Chase and Lev’s memory model. A good test-case is provided by the Chase–Lev deque running on top of a relaxed memory model. Furthermore, while work-stealing is widely used on the x86 architecture (an evaluation of the Cilk language for shared-memory multiprocessors [4], but thanks to its merits [2] it has been adopted in a number of parallel libraries and parallel programming environments, including the Intel TBB and compiler suite. Work-stealing variants have also been proposed for distributed clusters [5] and heterogeneous platforms [1]. The scheduling strategy is intuitive:

• Each processor uses a dynamic array as a doubly-ended queue (deque for short) holding tasks ready to be scheduled.
• Each processor manages its own deque as a stack. It may only push and pop tasks from the bottom of its own deque.
• Other processors may not push or pop from that deque; instead, they steal tasks from the top when their own deque is empty. In most implementations, the stolen deque is selected at random.
• Initially, one processor starts with the “root” task of the parallel program in its deque, and all other deques are empty.

The state-of-the-art algorithm for the work-stealing deque is Chase and Lev’s lock-free deque [3]. It uses an array with automatic, asynchronous growth. Assuming sequentially consistent memory, it involves only one atomic compare-and-swap (CAS) per steal, no CAS on push, and no CAS on take except when the deque has exactly only one element left.

We implemented and tested four versions of the concurrent deque algorithm, with different barrier configurations: (1) a sequentially consistent version, written with C11 seq. cst atomics, following the original description in [3]; (2) an optimized version, which takes full advantage of the C11 relaxed memory model, reported in Figure 1; (3) a native versions for ARMv7, reported in Figure 2,
and (4) a native version for x86. These native versions rely on compiler intrinsics and inline assembly to leverage architecture-specific assumptions and thus reduce the number of barriers required.

In our implementations of Figure 1 and Figure 2, we assume that the Deque type is declared as:

```c
typedef struct {
    atomic_size_t size;
    atomic_int buffer[];
    Atomic(Array *) array;
    Atomic(Array *) array;
} Deque;
```

In the code of Figure 1 the atomic and memory_order prefixes have been elided for clarity. The ARM pseudo-code of Figure 2 uses the keywords R and W to denote reads and writes to shared variables, and atomic indicates a block that will be executed atomically, implemented via LL/SC instructions. The x86 version is based on prior work [10] and only requires a single mfence memory barrier in `take`, in place of the call to `thread_fence` in the C11 code.

2.1 Notions of correctness

The expected behavior of the work-stealing deques is intuitive: tasks pushed into the deque are then either taken in reverse order by the same thread, or stolen by another thread. We say that an implementation is correct if it satisfies four criteria, formalized and proven correct for our ARM optimized code in Section 4:

1. tasks are taken in reverse order;
2. only tasks pushed are taken or stolen (well-defined reads);
3. a task pushed into a deque cannot be taken or stolen more than once (uniqueness);
4. given a finite number of push operations, all pushed values will eventually be either taken or stolen exactly once, if enough take and steal operations are attempted (existence).

These criteria hold because of the following assumptions and properties of the Chase–Lev algorithm:

- For any given deque, push and pop operations execute on a single thread. Concurrency can only occur between one execution of push or take in the owner thread, and one or more executions of steal in different threads.
- Newly pushed tasks are made visible to take and steal by the increment to `bottom` in push. As we shall see in Section 4, our ARM implementation enforces this by placing a sync barrier before the update of `bottom`, guaranteeing that the pushed element cannot be stolen before `bottom` is updated.
- Taken tasks are reserved first by updating `bottom`; again, in our ARM implementation, the sync barrier placed after the update to `bottom` will ensure that it will not be concurrently stolen.
- Stolen tasks are reserved by updating `top`. The only situation where `steal` and `take` contend for the same task is when the deque has a single element left; this particular conflict is resolved through the CAS instructions in both `take` and `steal`. This scenario was further optimized by Chase and Lev, making the CAS in `take` conditional upon the size of the deque being 1. The correctness of this optimization on a relaxed memory model depends on the presence of both full barriers in `take` and `steal`, to ensure that at least one of the participants will have a consistent view of the size of the deque. Having just one `take` or `steal` seeing a consistent view of the size of the deque is enough; if it is `take`, that will force a CAS to be performed; if it is `steal`, the index reservation will ensure an empty return value.
- Finally, stolen tasks are protected from being concurrently stolen multiple times by the monotonic CAS update to `top` in `steal`. This CAS orders `steal` operations and makes them mutually exclusive. At the same time, `steal` operations that abort due to a failed CAS do not change the state of the deque.

2.2 Comparison of the C11 and ARM implementations

Our C11 implementation in Figure 1 is optimal in the sense that no C11 synchronization can be removed without breaking the algorithm. However, if low-level atomics are compiled using the mapping of McKenney and Silvera [9] on ARM/POWER or the mapping of Tehrekov [14] on x86, the generated code contains more barriers than the hand-optimized native versions on both x86 and ARM. We show in Section 5 that this happens because of the need for seq-ctxt atomics to simulate ARM/POWER cumulative semantics. Concretely, on ARM, an extra dmb instruction is inserted before each CAS operation [11], compared to the native version where a relaxed CAS—coherent and atomic only—is sufficient. On x86, an mfence instruction is added between the two reads in `steal`. The fully sequentially consistent C11 implementation inserts many more redundant barriers [11].

3. The memory model of ARMv7

The memory model of the ARMv7 architecture follows closely that of the POWER architecture, allowing a wide range of relaxed behaviors to be observable to the programmer:

1. The hardware threads can each perform reads and writes out-of-order, or even speculatively. Basically any local reordering is allowed unless there is a data/control dependence or synchronization instruction preventing it.
2. The memory system does not guarantee that a write becomes visible to all other hardware threads at the same time point. Writes performed by one thread are propagated to (and become visible from) any other thread in an arbitrary order, unless synchronization instructions are used.
3. A dmb barrier instruction guarantees that all the writes which have been observed by the thread issuing the barrier instruction are propagated to all the other threads before the thread can continue. Observed writes include all writes previously issued by the thread itself, as well as any write propagated to it from another thread prior to the barrier. This semantics of barrier instructions is referred to as cumulative.

We build on the axiomatic formalization of POWER and ARMv7 memory model by Mador-Haim et al. [7], which has been proved equivalent to the operational semantics of Sarkar et al. [12]. A gentle introduction can be found in [8].

Axiomatic execution witnesses capture abstract memory events associated with memory-related instructions and internal transitions of the model. Unlike in stronger models such as x86, each memory access is represented at run-time by two distinct events: an issuing event—called sat for reads and ini for writes—eventually followed by a commit event when the speculative state of the instruction is resolved. Once a write instruction is committed, events that propagate it to other threads can be observed—propagation to thread A is denoted ppA. All the relations part of an execution witness are listed in Table 1.

The core of the axiomatic model builds on the evord relation, modeling the happens-before order between events. This satisfies the fundamental property:

```
word \rightarrow after \cup before \cup comm \cup mem \cup local
```

and must be acyclic for an execution to be consistent.

We assume that the atomic sections, used to represent CAS-like behaviors, are executed atomically and obey a total order. We model them either as a single instance of a read instruction (failed CAS) or an atomic read–write pair of instruction instances (successful CAS). The atomicity of these accesses is captured by the xatom relation. We do not assume any other property on these atomic sections (e.g., cumulativity). In practice, atomic sections can be implemented with LL/SC instructions.

We use several notation shortcuts. We refer to the deque global variables `top`, `bottom`, and `array` as `t`, `b`, and `a`. Elements of the buffer are written `x`, where `i` is the virtual index in natural numbers.
before any wrap-around is applied. Barrier instructions are omitted
to brevity when implied by the presence of a by or
relation. Irrelevant values in reads and writes are replaced with the
placeholder “” (e.g., R.c.). We do not label instruction instances
individually, but decorate them with a disambiguating execution
prefix, identified by a dot. These prefixes do not only distinguish
individually, but decorate them with a disambiguating execution
witness (hereafter, execution graphs) that cannot occur together in
the same consistence execution witness. We then show that all in-
correct executions, such as those containing two instances of steal
reading the same value added by a single instance of push, cannot
have consistent execution witnesses and, as such, cannot happen.

The proof is structured as follows. In 4.1 we provide basic tech-
nical definitions and properties of the memory model, which are
used throughout the proof. In 4.2 we describe all the possible exe-
cution graphs for each of the three operations (push, take or steal),
the control flow of the ARM code in Figure 2. In 4.3 we show
how the succession of dynamic arrays built by resizing can be
abstracted as a single sequence of unique abstract values indepen-
dent of resize operations, with strong coherence and consistency
properties. Corollary 2 establishes Criterion 2 (well-defined reads).
In 4.4 we build on the previous abstraction to prove Theorem 1,
pertaining to the uniqueness property of elements taken and stolen,
which corresponds to Criterion 3 (uniqueness). Finally, in 4.5, we
relate all previous results to prove Theorem 2 establishing Crite-
ron 4, i.e., the existence of matching

4. Proof of correctness of the ARM code

The proof is divided into five parts; it validates the criteria 2 to 4
equated in Section 2.1. Since push and take never execute con-
currently and b is only ever modified in one of these functions,
the proof of Criterion 1 does not involve reasoning about concurrency
and is omitted here.

The proof builds on a precise analysis of all the possible execu-
tion witnesses of arbitrary invocations of the algorithm. We recall
that an execution witness, as defined by the ARM axiomatic model,
is a graph capturing all memory events occurring during an execu-
tion (vertices), as well as the relations that link them (edges). Indi-
vidual lemmas strive to narrow down the set of possible execution
witnesses, based on properties of the algorithm and of the architec-
ture. To that end, we pinpoint specific subgraphs of an execution

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.pdf}
\caption{C11 code of Chase–Lev deque, with low-level atomics}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.pdf}
\caption{ARM pseudo-code of Chase–Lev deque}
\end{figure}
**Table 1.** Summary of relations used in the ARM axiomatic model


1. Note that \( \text{pp-sat} \) does not imply an event happens-before order on the events making up the related instruction instances.
4.2 Execution paths

We consider the three operations of the work-stealing algorithm: take, push and steal. Each of them exhibits different execution paths depending on control flow. Data and address dependences are implicit in the notations and are omitted for brevity. Control dependences are implied by the guard conditions in each case and are also omitted, but we explicit the constraints on the b and t variables carrying the control dependence. Greek letters $\beta$, $\tau$, $\xi$ denote the memory values of $b$, $t$, and some $x_i$, respectively. Reads and writes are annotated with the corresponding line from Figure 2.

For take and steal, we say that an instance of the operation is successful if it returns one element; otherwise (including if it returns empty) it is considered failed.

4.2.1 Take

Two failure cases return no element (empty), and two success cases return one element from the deque. All four paths start with:

$$ (a) Rb, 0, \beta R, x, c W b, x, \beta - 1, \xi W, d R t, \tau $$

Specific continuations for each path are listed below.

**Return empty without CAS**, $\beta - \tau \leq 0$:

$$ \ldots \xi W, d R t, \tau $$

**Return empty with (failed) CAS**, $\beta - \tau - 1, \tau \neq \tau'$:

$$ \ldots \xi W, d R t, \tau $$

**Return one without CAS**, $\beta - \tau > 1$:

$$ \ldots \xi W, d R t, \tau $$

**Return one with (successful) CAS**, $\beta - \tau = 1$:

$$ \ldots \xi W, d R t, \tau $$

4.2.2 Push

There are two paths: a straight case, and a resizing case which grows the underlying circular buffer.

**Straight**, $\beta - \tau < \text{size}(x) - 1$:

$$ (a) Rb, \beta R, c R, x, c W b, x, \beta + 1 $$

**Resizing**, $\beta - \tau \geq \text{size}(x) - 1$:

$$ (a) Rb, \beta R, c R, x, c W b, x, \beta + 1 $$

4.2.3 Steal

There are three paths: two failure cases and one success case. Failure returns no element but returns a failed element.

**Return empty without CAS**, $\beta - \tau \leq 0$:

$$ (a) Rb, 0, \beta R, x, c W b, x, \beta - 1, \xi W, d R t, \tau $$

**Return empty with (failed) CAS**, $\beta - \tau > 0, \tau \neq \tau'$:

$$ (a) Rb, 0, \beta R, x, c W b, x, \beta - 1, \xi W, d R t, \tau $$

**Return one with (successful) CAS**, $\beta - \tau > 0$:

$$ (a) Rb, 0, \beta R, x, c W b, x, \beta - 1, \xi W, d R t, \tau $$

4.3 Significant reads and writes

We define the sequence $(\beta_n)$ of values taken by the variable $b$ over the course of the program, according to the write coherence relation. Initially $\beta_0 = 0$. Since all push and take operations occur in a single thread, and steal operations never alter the value of $b$, the elements of $(\beta_n)$ correspond to writes to $b$ in program order within the push and take operations. Similarly, we define the sequence $(\tau_n)$ of values taken by the variable $t$. We assume $\tau_0 = 0$.

Furthermore, since all writes to $t$ are from CAS instructions, which are sequentially ordered, and all such CAS instructions increment $b$ by one, $(\tau_n)$ is monotonically increasing, and s.t. $\tau_n \geq \tau_{n-1}$. For each index $i$, we define the sequence $(\xi_i)_{i \in \epsilon_i}$ of successive values given to the element at index $i$ in the deque by the last write $x_i$. of a push operation, regardless of the address $\&x$ of the underlying array. Only the last such write is called significant as it induces a new value in an $(\xi_i)$ sequence, while writes due to resizing do not. For all $i$, $\xi_i$, the value before the first significant write to $x_i$ location, is undefined: $\xi_0 = \perp$. Similarly, a read is significant if it occurs in a successful instance of take or steal.

**Lemma 4.** For all $i$, $(\xi_i)$ is globally coherent.

**Proof.** Given two significant writes $W_{x_i}$ and $W_{x_i}'$, at index $i$ (regardless of the address of the underlying array). If $W_{x_i}$ and $W_{x_i}'$ both write to the same memory location, then they are ordered by write coherence. If they do not, then there must be a resize operation after the first write and before the second (all writes happen in the same thread). Because of the cumulative barrier after a resize operation, threads that see the second value must have seen the first beforehand. Hence, there is a global coherence order on the writes, which corresponds to the order of push operations.

We define the relation read from as follows: for some memory locations $m_0, \ldots, m_n$ and some value $v$, $W_{m_0}, v, \xi_0 R, m_0, v$ if $W_{m_0}, v \& R, m_0, v$ or there exists a sequence of copies carrying the value of the write to the read:

$$ W_{m_0}, v \& R, m_0, W_{m_1}, v \& R, \ldots \& W_{m_n}, v \& R, m_0, v $$

For conciseness, we hereafter omit the variable name from reads and writes whenever the variable can be inferred from the value: e.g., $W_{b_0}$ stands for $W_{b_0}, v$. Let $W_{\xi_i}$ denote the $i$th significant write at index $i$, and $R_{\xi_i}$ a significant read s.t. $W_{\xi_i} \& R, R_{\xi_i}$.

**Lemma 5.** Given a write $W_{x_i}$, and a read $R_{x_i}$.

$$ i \neq j \implies W_{x_i} \& R_{x_i} $$

**Proof.** If the addresses of the underlying arrays differ, then the memory locations read and written are distinct and there can be no read from relation.

Otherwise, since old arrays are never reused, the addresses are the same and $i \equiv j \mod \text{size}(x) R_{x_i}$, belongs to a successful instance of take, push (with resizing), or steal. Let $X$ be that instance.

Let $P$ be the instance of push to which $W_{x_i}$, belongs. In $P$, we have the following execution graph:

$$ P R_t, \tau P, c R_{x_i}, \xi W_{x_i}, R, v, \beta + 1 $$

where $\tau P \leq i \leq \beta P$ and $\beta P - \tau P < \text{size}(x) - 1$

Let us assume $i \neq j \land W_{x_i}, c R_{x_i}$, and show it is indeed impossible.

Assume $X$ is a successful instance of take or push. Since $X$ and $P$ belong to the same thread, $P$ must occur before $X$ in program order (the order of loads and stores to the same location is preserved: $P W_{x_i} \& R_{x_i})$.

If $j < i$, then $j \leq i - \text{size}(x)$. However, the following must hold in $P$:

$$ \tau P \leq i \leq \beta P - \tau P < \text{size}(x) - 1 $$

hence $j < i - \text{size}(x) + 1 \leq \beta P - \tau P < \text{size}(x) - 1 < \tau P$

Furthermore, if $X$ is a take operation, $R_{x_j}$ reads the last element of the deque, and $j = \beta X - 1 \geq \tau X$, if $X$ is a push operation, $R_{x_i}'$, results from a copy operation of the resizing code, hence $j \geq \tau X$. Since $X$ occurs after $P$ in program order and $t$ is monotonically increasing, $P R_t, \tau P, c R_{x_i}, \xi W_{x_i}, R, v, \beta P - \tau P < \tau X$. If $i < j$, then, since $i \geq \beta X$, $b$ must increase from $\beta P + 1$ to $j + 1$ between the write in $P$ and the read in $X$. Hence, there must be an instance $P'$ of push between $P$ and $X$ (in program order) that increments $b$ to $j + 1$. Indeed, the only writes that increase the value of $b$ occur in push and take; and the effect of take as a whole never increases the value of $b$ since it first decrements the variable. We have:

$$ P W_{x_i}, c R_{x_i}, \xi W_{x_i}, R_{x_j}, \xi X, R_{x_j} $$

hence $P W_{x_i}, c R_{x_i}, \xi W_{x_i}, R_{x_j}, \xi X, R_{x_j}$.
Thus, from Lemma 2 (ii), $P \cdot W x_i, x_j \rightarrow_\beta R z_j$. 

Now, assume $X$ is a successful instance of $steal$. We have the following execution graph for $X$: 

$$X \cdot R t, x, \tau_X = \begin{cases} \text{if } i < j \text{ then } x \cdot \text{size}(x) & \text{if } j = i \text{ then } x \cdot \text{size}(x) + 1 \end{cases} \cdot \text{size}(x) + 1$$ 

If $i < j$, then $j = i \cdot \text{size}(x)$. However, the following must hold in $P$: 

$$i < j \cdot \text{size}(x) + 1 \leq \beta \cdot \text{size}(x) + 1 < 1 + \tau_p$$

Hence $\tau_X = j \cdot \tau_p$. Since $t$ increases monotonically, it must be that: 

$$X \cdot R x_j \rightarrow_\beta \text{size}(x) \cdot R t, x, \tau_p \cdot \text{size}(x) + 1$$

Hence $X \cdot R x_j$ must be committed before $R t, x, \tau_p$. Since $R t, x, \tau_p + 1$ is (cumulatively) propagated to $W x_i, x \cdot R x_j$, $X \cdot R x_j$ must be committed before $W x_i$. Formally: it follows from Lemma 3 (ii) that $W t, x, \tau_p + 1 \cdot \text{size}(x)$, $R t, x, \tau_p$. If $W x_i, x \cdot \text{size}(x) \rightarrow_\beta R z_j$, then $W x_i, x \cdot \text{size}(x) \rightarrow_\beta R z_j$. We get: 

$$X \cdot W t, x, \tau_p + 1 \cdot \text{size}(x), x \cdot R x_j \rightarrow_\beta \text{size}(x) \cdot R t, x, \tau_p + 1$$

Lemma 2 (ii) tells that it is impossible. Thus $P \cdot W x_i, x_j \rightarrow_\beta X \cdot R x_j$. 

Otherwise, we would have $P \cdot W x_i, x_j \rightarrow_\beta X \cdot R x_j$. 

Let $\xi = \exists v \cdot \text{size}(x)$, $\xi = \text{size}(x)$, $\nu = \text{size}(x) + 1$. Since $\nu > \xi$, it must be that: 

$$W x_i, x_j \rightarrow_\beta x \cdot \text{size}(x) \cdot R x_j$$

It follows from Lemmas 2 (iii), 2 (ii) and 2 (i) that $W x_i, x_j \rightarrow_\beta R x_j$. 

Hence, Lemma 2 (ii), $W x_i, x_j \rightarrow_\beta R x_j$. 

Corollary 1. Given a significant write $W x_i$ and a significant read $R x_j$, $i \neq j \Rightarrow W x_i \rightarrow_\beta R x_j$. 

Proof. If $i = j$, we know that $W x_i \rightarrow_\beta R x_j$. Furthermore, all copies, which happen during a resize operation, copy from and to the same index. Since there are less copies than the size of the expanded array, there can be no two copies writing to the same memory location in the new array. Hence, there can be no sequence of copies from $W x_i$ to $R x_j$. 

Lemma 6. Given a significant write $W x_i$ and a significant read $R x_j$: 

(i) $W x_i \rightarrow_\beta R x_j, x \cdot \text{size}(x) \Rightarrow u \leq v$ 

(ii) $0 < u \leq v \Rightarrow W x_i \rightarrow_\beta R x_j, x \cdot \text{size}(x)$ 

Proof. We prove each point separately: 

(i) Suppose $v = u$. We define $W \cdot W x_i, x_i \rightarrow_\beta x_i$ as follows. 

If $u = 0$, $x_i$ is an undefined value; let $W \cdot W x_i, x_i \rightarrow_\beta x_i$. If $u < 0$, $x_i$ is the value of $x_i$. 

Otherwise, $0 < u < v$. Let $W \cdot W x_i, x_i \rightarrow_\beta x_i$. In other words, there exists a sequence of copies carrying the value of $x_i$ to $R x_i$. That sequence ends with a write $W \cdot W x_i, x_i \rightarrow_\beta x_i$. 

Moreover, according to the definition of $x_i$ and the semantics of resizing, $W \cdot W x_i, x_i \rightarrow_\beta x_i$. 

We have two cases: either $W x_i$ and $R x_i$ refer to the same memory location or they do not. 

Assume that they refer to the same memory location $x_i$. Then it must be that 

$$W \cdot W x_i, x_i \rightarrow_\beta x_i, x_i \rightarrow_\beta R x_i$$

and we have: 

$$W \cdot W x_i, x_i \rightarrow_\beta x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$$

Hence, from Lemma 2 (ii), $W \cdot W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. 

Conversely, assume that they do not refer to the same memory location. 

Then there must be a resize operation between $W \cdot W x_i, x_i \rightarrow_\beta R x_i$, and $W \cdot W x_i, x_i \rightarrow_\beta R x_i$: 

$$W a, x_i \rightarrow_\beta W a, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$$

Hence, from Lemma 3 (b), $W a, x_i \rightarrow_\beta W a, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. 

And from Lemma 2 (ii), $W a, x_i \rightarrow_\beta W a, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. 

Since there is only one write $W a, x_i$ that gives the value $x_i$ to $a$, we have a contradiction. 

(ii) There exists a write $W \cdot W x_i, x_i \rightarrow_\beta W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$, and a sequence of copies carrying the value of $x_i$ to $R x_i$. 

That sequence ends with a write $W \cdot W x_i, x_i \rightarrow_\beta W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. 

Therefore, $W x_i, x_i \rightarrow_\beta W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. 

Formally, it follows from Lemma 3 (ii) that $W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. We get: 

$$W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$$ 

Thus, we have $W x_i, x_i \rightarrow_\beta R x_i, x_i \rightarrow_\beta R x_i$. 

Theorem 1. Given a significant write $W x_i$ and a significant read $R x_j$, $i \neq j \Rightarrow W x_i \rightarrow_\beta R x_j$. 

Proof. If $i = j$, we know that $W x_i \rightarrow_\beta R x_j$. Furthermore, all copies, which happen during a resize operation, copy from and to the same index. Since there are less copies than the size of the expanded array, there can be no two copies writing to the same memory location in the new array. Hence, there can be no sequence of copies from $W x_i$ to $R x_j$. 

4.4 Uniqueness of significant reads 

The results from the previous section establish that two significant reads at different indexes cannot retrieve the same element $x_i$. The only possible cause of duplicate significant reads is thus reduced to the case where the reads access the same index $i$. 

Theorem 2. Given a significant write $W x_i$ and a significant read $R x_j$, $i \neq j \Rightarrow W x_i \rightarrow_\beta R x_j$. 

Proof. If $i = j$, we know that $W x_i \rightarrow_\beta R x_j$. Furthermore, all copies, which happen during a resize operation, copy from and to the same index. Since there are less copies than the size of the expanded array, there can be no two copies writing to the same memory location in the new array. Hence, there can be no sequence of copies from $W x_i$ to $R x_j$. 

Lemma 7. Given $S_1$ and $S_2$ distinct successful instances of steal, 

$$\forall x_i \in S_1, \exists x_i \in S_2, x \neq x_i \cap \neq x_i \neq x_i$$ 

Proof. All writes to $t$ atomically increment it (by atomicity of CAS). Hence two successful steal operations cannot write (thus read) the same value of $t$. Reads from $x$ in a steal operation access the index given by the value of the $t$ variable. Hence, if $i \in S_1$ and $t' \in S_2$ imply it $i \neq i'$. 

Lemma 8. Given $T$ a successful instance of $take$ and $P$ an instance of $push$. If $P$ comes after $T$ in program order, then: 

$$\forall x_i \in T, \forall x_i \in P, x \neq x_i \cap x \neq x_i$$
Proof. Assume $i = j \land v = u$. We have $R_{\xi_i}^{\ast} \rightarrow^{\ast} W_{\xi_j}^{\ast}$, therefore $W_{\xi_i}^{\ast} \rightarrow^{\ast} R_{\xi_j}^{\ast}$. From Lemma 6 (ii), it follows that $u > v$. We have a contradiction. \hfill $\Box$

Lemma 9. Given $T_1$ and $T_2$ distinct successful instances of take, $\forall R_{\xi_i}^{\ast} \in T_1, \forall R_{\xi_j}^{\ast} \in T_2, i \neq i' \lor v \neq v'$

Proof. We have the following execution graphs:

$T_1: R_{\beta_1} \rightarrow^{\ast} W_{\beta_1} b, \beta_1 - 1 \rightarrow^{\ast} R_{\tau_1} \rightarrow^{\ast} R_{\beta_1 - 1} \rightarrow^{\ast} \cdots$

$T_2: R_{\beta_1'} \rightarrow^{\ast} W_{\beta_1'} b, \beta_1' - 1 \rightarrow^{\ast} R_{\tau_1'} \rightarrow^{\ast} R_{\beta_1'-1} \rightarrow^{\ast} \cdots$

And $\beta_1 - 1 = i$ and $\beta_1' - 1 = i'$. Since all instances of take occur in the worker thread, we have either: $T_1: W_{\beta_1} b, \beta_1 - 1 \rightarrow^{\ast} R_{\tau_1} \rightarrow^{\ast} T_2: R_{\beta_1'}$. Let us assume the first case as well as $i = i' \land v = v'$ and show it is impossible, the other case being symmetrical. We have $\beta_1 - 1 = i = i' = \beta_1' - 1$, and $T_1: W_{\beta_1} b, \beta_1 - 1 \rightarrow^{\ast} R_{\tau_1} \rightarrow^{\ast} T_2: R_{\beta_1'}$. Hence ($\beta_1$) must increase from $i$ to $i + 1$, between $n$ and $n'$; there exists an instance $P$ of push that writes $W_{\beta_1} \rightarrow^{\ast} \cdots T_2: R_{\beta_1'}, b, \beta_1' - 1 \rightarrow^{\ast} \cdots$ and $\beta_1 - 1 = i$ and $\beta_1 = i + 1$ (as noted above, take as a whole does not increase the value of $b$). We get the following graph:

$W_{\beta_1} b, \beta_1 - 1 \rightarrow^{\ast} T_2: R_{\beta_1'} \rightarrow^{\ast} R_{\beta_1} \rightarrow^{\ast} \cdots$ \hfill $\Box$

Lemma 10. Given $T$ a successful instance of take and $S$ a successful instance of steal, $\forall R_{\xi_i}^{\ast} \in T, \forall R_{\xi_j}^{\ast} \in S, i \neq i' \lor v \neq v'$

Proof. We have the following execution graphs:

$T: R_{\beta_1} \rightarrow^{\ast} R_{\beta_1 - 1} \rightarrow^{\ast} W_{\beta_1} b, \beta_1 - 1 \rightarrow^{\ast} W_{\beta_1} \rightarrow^{\ast} \cdots S.R_{\tau_m}, \rightarrow^{\ast} S.R_{\tau_m}, \rightarrow^{\ast} \cdots$ \hfill $\Box$

Lemma 11. Given $X$ a successful stealing instance of take or steal: $\forall R_{\xi_i}^{\ast} \in X, X, m > m'$. Proof. We have two cases:

$\bullet$ Assume $X$ is an instance of take. $X$ follows $P_{F_r} \rightarrow^{\ast} X.R_{\tau_m}, m > m$ by uniprocessor constraints.

$\bullet$ Assume $X$ is an instance of steal. Since $X$ is successful, $X$ contains a successful instance of a CAS instruction, hence the two reads from $t$ must yield the same value. Due to the barrier between $X.R_{\tau_m}$ and the following read $X.R_{\tau_m}$, and the barrier before $P_F.W_{\beta_m} \rightarrow^{\ast}$, we have: $W_{\tau_m} \rightarrow^{\ast} \cdots W_{\beta_m} \rightarrow^{\ast} X.R_{\beta_m} \rightarrow^{\ast} W_{\beta_m}$. From Lemma 3 (ii), we have $W_{\tau_m} \rightarrow^{\ast} X.R_{\beta_m} \rightarrow^{\ast} X.R_{\tau_m}$. It then follows from Lemma 3 (i) that $W_{\tau_m} \rightarrow^{\ast} X.R_{\beta_m} \rightarrow^{\ast} X.R_{\tau_m}$. Total order on CAS instructions and Lemma 2 (iii) guarantee that $X.R_{\beta_m} \rightarrow^{\ast} X.R_{\tau_m}$ where $m \geq m'$. Therefore, $m \geq m'$. \hfill $\Box$
Let $\beta_t \leq \tau_m$ or because of a failed CAS instruction. Suppose there is no $X$ where $\beta_t \leq \tau_n$; then all failures must be due to a failed CAS instruction. A failed CAS occurs if the two values of $t$ read during the instance $X$ differ. Let $Y_1$ and $Y_2$ be two such failed instances executing in a same thread; let us assume that $Y_2$ follows $Y_1$ in program order, $n_1 \neq n'_1$ and $n_2 \neq n'_2$. There exists a write $W_{n_1} \rightarrow_{\text{po-lo c}} R_{n_1} \rightarrow_{\text{po-lo c}} Y_2 \rightarrow_{\text{po-lo c}} R_{n_2} \rightarrow_{\text{po-lo c}} R_{n'_2}$.

One may finally prove Theorem 2. On the one hand, Corollary 4 tells that the number of significant reads (from a successful instance of $\text{take or steal}$) is equal to the number of significant writes (from an instance of $\text{push}$). On the other hand, Theorem 1 states that significant reads uniquely map to significant writes. By injectivity, there exists a unique significant read for each significant write.

5. On the C11 implementation

The configuration consistent implementation is a direct translation of the original algorithm using C11 $\text{seq_cst}$ atomic variables for all shared accesses. It is obtained from the code in Figure 1 by replacing all memory order constants with $\text{seq_cst}$; doing so makes fences unnecessary, hence they should be removed.

The optimized C11 implementation improves upon the previous version by replacing sequential consistency with release–acquire pairs where appropriate. It establishes happens-before relations between reads and writes, as required by the proof. Unfortunately, without relying on $\text{seq_cst}$, using only release, acquire and consume operations, we were unable to reproduce the required memory ordering constraints needed on the POWER and ARM architectures while adhering to C11 semantics.

Although designed for POWER/ARM, most of the arguments developed in the proof in turn transfrom to the rules of C11 in a straightforward fashion. In all cases that do not involve cumulativity, the $\text{seq_cst}$ relation (defined in 4.1) combined with dependences, which form the core of the POWER/ARM proof, may be replaced with analogous properties pertaining to the C11 happens-before relation combined with release–acquire semantics. The one notable difference between the two models lies in the absence of cumulativity in the design of the C11 abstract machine: neither C11 fences nor C11 atomic accesses guarantee cumulativity. A similar effect can be achieved by chaining alternating release–acquire writes and reads, which form a happens-before path. But this device does not work in situations where propagation needs to be asserted between two reads, rather than a read followed by a write. This situation occurs in the steal operation. Informally (see Lemma 10 for the formal description), it must be that two concurrent steal and take do not read “old” values of both bottom and top, where “old” could be defined as “older than the value known to the other party in coherence order”. The presence of the two cumulative barriers in $\text{steal}$ and $\text{take}$ on ARM guarantees such a condition:

- if the $\text{take}$ barrier is ordered before the barrier in $\text{steal}$, then the program-order-previous write to bottom will be propagated to the instance of $\text{steal}$;
- conversely, if the $\text{steal}$ barrier is ordered before the barrier in $\text{take}$, then value read by the program-order-previous read from top will be propagated to the instance of $\text{take}$.

In the second case, it is important to remark that the write that produced the value read in $\text{steal}$ might belong to another thread, and thus not be sequenced before the barrier. In the absence of cumulativity, it need not be propagated to the instance of $\text{take}$.

To enforce this particular case of cumulativity in C11, we rely on the properties of sequential consistency. By making all writes (actually, CAS operations) to top sequentially consistent, we ensure that there is a total ordering between the two fences (in $\text{take}$ and $\text{steal}$) and the write that produced the value of top read in the instance of $\text{steal}$. Furthermore, if that read uses acquire semantics, then there is a happens-before relation between it and the $\text{steal}$ barrier. Hence, the write must come before said barrier in sequential consistency total order. Then, either the barrier in $\text{steal}$ is ordered before the barrier in $\text{take}$, or the other way around:

- if the $\text{steal}$ barrier is ordered before the barrier in $\text{take}$, then it follows from $\text{seq_cst}$ barrier semantics that the value of top read by $\text{take}$ cannot be older than the one read in $\text{steal}$;
- conversely, if the $\text{take}$ barrier is ordered before the one in $\text{steal}$, then the value of bottom read by $\text{steal}$ cannot be older than the one written in $\text{take}$.

6. Experimental results

We present experimental results on three current and widely used architectures: (1) a Tegra 3 ARMv7 processor rev 9 (v7l) with 4 cores at 1.3GHz and 1GB of RAM; (2) an Intel Core i7-2720QM machine with 4 cores (hyper-threading disabled) at 2.2GHz and 4GB of RAM; and (3) a dual-socket AMD Opteron Magny-Cours 6164HE machine with 2 × 12 cores at 1.7GHz and 16GB of RAM. All tests were compiled with GCC 4.7.0, the first release of GCC to introduce built-in support for C11 atomics.

6.1 Synthetic benchmarks

We designed a synthetic benchmark to simulate the depth-first traversal of a balanced tree—with breadth $b$ and depth $d$—of empty tasks by a main worker thread, reproducing the prototypical execution of a Cilk program. One or more thieves attempt to steal these tasks. For robustness and predictability, the worker always creates
Figure 3. Synthetic single-thief benchmarks

Figure 4. Synthetic multi-thief benchmarks

and pushes the same number of tasks in the deque, following the depth-first pattern, regardless of whether a specific continuation has been stolen by another thread (it is simply recorded as stolen, but subsequent tasks spawn normally and locally). The thieves perform steal actions at a configurable rate, and discard stolen tasks.

We have experimented with two different methods of steal distribution, the goal being to uniformly spread the contention over the entire life of the worker thread. The first method is based on the CPU clock of the core dedicated to each thief; with this technique, the clock is regularly sampled and the appropriate number of steal operations is performed accordingly. The second method relies on a random number generator, called in a busy loop, which allows steal operations with a set probability. While the clock-based approach produces more reliable results, it can only be used if a low-overhead CPU clock is available from user space, which is unfortunately not the case on our ARM-based system.\(^6\) Conversely, the second technique suffers from imprecision when targeting smaller ranges of frequencies, which is necessary on faster processors or when the number of cores increases.\(^7\) Hence, the former is used on x86 and the latter on ARM, with appropriate empirical tuning to gather results over a common representative range of steal throughput.

We selected two workloads: a reasonably broad tree \((b = 3; d = 15)\) and, as a special case, a degenerate comb-shaped tree \((b = 1; d = 10^7)\). The former is meant to reproduce normal contention with steal operations alongside both push and take, while the latter illustrates a case of contention between take and steal only.

\(^6\) The ARM C15 cycle counter register can only be queried if first enabled from kernel mode, and is delegated to a monitoring co-processor, with unclear consequences for the bus, caches, and memory model as a whole.

\(^7\) On higher end processors with multiple cores acting as thieves, higher steal probabilities can yield many times more steal attempts than there are tasks created over a set period.
We measure the time taken by the worker thread to complete the specified number of task creations and consumptions. This in turn serves to compute the pushlake throughput—the combined number of push and take operations completed per unit of time, as well as the effective steal throughput, defined as follows: the test protocol strives to perform a number of steals over time, at a fixed, nominal steal throughput; the effective throughput is the real throughput as could be observed after the experiment, i.e., how many steals were actually performed during the lifetime of the worker thread. These metrics provide a measure of the efficiency of the algorithm on its critical path at various levels of contention.

In order to assess the impact of the added barriers on the different architectures, raw throughput values have been normalized by the near-ideal throughput on the same workload (see Table 2), obtained on a single thread with no contention and no synchronization: memory barriers are replaced with simple compiler fences, and CAS operations with a simple branch and conditional assignment. These numbers provide a good approximation of the upper bound on the achievable throughput on each machine, though other minor factors can contribute to higher observable values. In particular, it should be noted that counting the throughput in number of operations per second is, by design, a generalization: the execution time of each operation depends on its nature and the exact control path taken; for example, an invocation of take returning empty will be faster than one returning a task.

In all diagrams, we have included a set of points labeled nofences, for comparison purposes. These correspond to the least common denominator among all the tested barrier placement strategies: only relaxed CAS operations are included, with otherwise no memory barriers. The nofences version violates the semantics of the work-stealing deque. Each of our proposed implementations of the algorithm can be seen as adding a different set of barriers to nofences, making it correct. Hence, results obtained with nofences should be taken as no more than a general baseline, as the complete lack of fences can lead to unexpected behavior. For instance, Figures 3 and 4 show greater throughput values at high contention for the comb-shaped workload on ARM. Those are the results of a long tail of fast empty take operations, an artifact due to the nature of the comb-shaped test and the absence of synchronization between empty take and steal (enabled by the lack of barriers in take).\(^8\) (peak at above 50%), indicating that the first serializing instruction introduced in the code is very costly on x86, especially if it is added to the critical path (as is the case in native, c11 and seqcst, but not in nofences). This could suggest either the stronger guarantees of the x86 memory model—a full memory fence is required to linearize history in order to maintain total store order [13]—or aggressive local optimizations for single-thread execution without communication.

From these observations, we can postulate that advanced ARM architectures such as the Tegra 3 benefit the most from a well-written concurrent program that takes full advantage of the flexibility allowed by their memory model, and conversely struggle relatively more with literal interpretations of algorithms designed with stricter, simpler hypotheses in mind.

### 6.2 Task-parallel benchmarks

We further experiment on common task-parallel benchmarks, mostly extracted from the Cilk benchmark suite,\(^9\) to evaluate the impact of the memory barrier optimization on realistic workloads and load-balancing scenarios.

**Fibonacci** is the tree-recursive computation of the 35\(^\text{th}\) Fibonacci number; it illustrates the raw cost of the scheduling algorithm as each task only performs a single addition.

**FFT-1D** computes the Cooley-Tukey fast Fourier transform over a vector of 2\(^{20}\) elements.

**Matmul** is the blocked matrix multiplication, of size 256\(\times\)256 on the Tegra 3 and Core i7 platforms, and of size 384\(\times\)384 on Opteron to ensure a sufficient computation time.

**Strassen** is an optimized matrix multiplication algorithm, running on matrices of size 512\(\times\)512 on the Tegra 3 and Core i7 platforms, and of size 2048\(\times\)2048 on Opteron.

**Knapsack** is the usual resource allocation problem. A set of objects, each with a given weight and value, must be picked from a pool to fit a total weight constraint while maximizing value. We use 33 objects.

**Seidel** simulates heat transfer using the Gauss-Seidel method which iterates a 5-point stencil over an two-dimensional array. We used a resolution of 1024\(\times\)1024 points with 20 iterations.

We compare the four implementations of the Chase–Lev deque presented in Section 2. The sequentially consistent, direct translation to C11 serves as a baseline to measure the speedups obtained with the three other implementations. We observe that the nofences version is inherently incorrect and generally results in erroneous executions. This version is only presented as a rough upper-bound on the performance gains of memory barrier optimization.

Figure 5 shows similar trends to the balanced tree synthetic kernel, with a clear advantage to the two optimized implementations (native and c11) over the seqcst baseline. **Fibonacci**, **FFT-1D**, **Matmul** and **Strassen** use a recursive divide-and-conquer pattern, leading to balanced binary trees. They appear in increasing order of granularity, ranging from a single addition to the multiplication of matrix blocks of size 16\(\times\)16. On **Fibonacci**, the lowest granularity kernel, the results are very similar to the throughput achieved by the synthetic benchmark: the optimized versions show up to 1.19\(\times\) speedup on the Tegra 3 platform against the seqcst version, 1.3\(\times\) on Core i7 and up to 1.13\(\times\) on Opteron. These speedups gradually decrease as granularity increases, hiding the cost of the scheduling deque. Yet we still observe significant speedups on the Matmul kernel, with a granularity of 64 (vector) multiply-add operations per task: we obtain up to 1.03\(\times\) speedup on Tegra 3,\(^10\) 1.1\(\times\) on Core

---

\(^8\)In the case where the deque is empty, neither take nor steal needs to execute a CAS instruction; furthermore, in the absence of barriers, the ARM memory model does not require successive decrements and increments of bottom in take to propagate to the thieves.

\(^9\)http://supertech.csail.mit.edu/cilk

\(^{10}\)The 0.95\(\times\) slowdown for the native version is a compiler artifact related with the usage of inline assembly.
i7 and 1.04× on Opteron. On Strassen, the largest granularity kernel, we no longer observe any significant improvement: the deque operations are entirely hidden by the work performed in each task. The Knapsack kernel is also based on a divide-and-conquer pattern, yet it does not result in a balanced tree because of the non-deterministic, dynamic pruning of sub-optimal branches. Unsynchronized communications are used to share the best total value reached on any branch; this value is used to stop exploring branches known to represent sub-optimal prefixes. The nofences version shows lower performance, on Tegra 3, than our optimized c11 and native versions because of longer delays until the best value is propagated to all cores, resulting in less pruning and more work. The performance improvement is reduced on this benchmark because additional barriers improve the accuracy on the current best value.

Finally, the Seidel kernel iterates on skewed wave-fronts of data-parallel tasks. A single main thread is responsible for spawning every task in a wave-front. It en-queues all tasks on its own work-stealing deque until it reaches a synchronization barrier. This scheme relies on stealing exclusively for the distribution of work among cores. This behavior puts a lot of strain on a particular deque, and induces a high level of contention. This explains the high performance gains of our optimized implementations on the 24-core Opteron platform: up to 1.3× speedup against the seqcst baseline, even higher than the speedups observed at a lower granularity on Fibonacci. However, despite its somewhat low granularity, corresponding to 16 additions and 4 multiplications of double precision floating point values per task, this benchmark only shows up to 1.05× improvement from our optimized versions on the Tegra 3 platform and up to 1.2× speedup on the Core i7. This is in line with the equivalent speedups observed for the similar granularity on FFT-1D, as the low number of cores on these platforms induce much less contention compared to the Opteron configuration.

Interestingly, our experiments show that the performance of our optimized versions is very close to the nofences version on the Tegra 3 platform. This validates our hypothesis about the benefits of an implementation that takes full advantage of the ARM relaxed memory model, rather than translating the classical sequentially consistent algorithm. Furthermore, the large performance gains on the two x86 platforms show that even in the case of stricter memory models such as total store order [13], relying on sequentially consistent algorithms represents a significant source of overhead.

7. Conclusion

We provided optimized implementations of Chase and Lev’s concurrent deque, targeting the weak memory models of the POWER and ARMv7 processors, as well as the C11 standard. Based on recent progress in the formalization of memory consistency, we established the first proof of the Chase–Lev deque for the weak memory model of a real-world processor. This is an important achievement, paving the way for robust parallel library and programming language implementations based on a work-stealing scheduler.

We obtained strong performance gains on ARM and x86 platforms, comparing our optimized implementation with portable C11 versions. For typical benchmarks, we showed that careful optimizations reach performance levels comparable to an (incorrect) fence-free version on a multicore ARM. This indicates that a high-throughput scheduler can be implemented efficiently on a weak memory model, benefiting from its scalability and energy savings.

Acknowledgments This work was partly supported by the European FP7 projects PHARAON id. 288307 and TERAFLUX id. 249013, and by ANR grant ANR-11-JS02-011.

References