

The MultiQueue

A Simple and Fast Relaxed Concurrent Priority Queue

MARVIN WILLIAMS, Karlsruhe Institute of Technology, Germany

PETER SANDERS, Karlsruhe Institute of Technology, Germany

Priority queues are used in a wide range of applications, including prioritized online scheduling, discrete event simulation, and greedy algorithms. In parallel settings, classical priority queues often become a severe bottleneck, resulting in low throughput. Consequently, there has been significant interest in concurrent priority queues with relaxed semantics.

In this article, we present the *MultiQueue*, a highly scalable and flexible relaxed priority queue. Its design is based on leveraging multiple internal priority queues using the seemingly paradoxical technique of “wait-free locking”. The practical performance of the MultiQueue is further enhanced by element buffering, batched operations on the internal queues, and cache-optimized access patterns.

We evaluate the quality-throughput trade-off of the MultiQueue against state-of-the-art competitors. The quality is measured using two complementary metrics: *rank error* (the distance of a deleted element to the best element), and *delay* (the number of elements with lower priority deleted before a given element). Extensive experiments on both micro-benchmarks and benchmarks based on real-world applications (shortest-path and branch-and-bound) show that the MultiQueue consistently outperforms its competitors. Additionally, its design allows users to easily tune the balance between throughput and quality to meet specific application requirements, and we believe that “wait-free locking” may be of broader interest for converting sequential data structures into high-performance, relaxed concurrent ones.

CCS Concepts: • **Computing methodologies** → **Concurrent computing methodologies**.

Additional Key Words and Phrases: concurrent data structure, priority queues, randomized algorithms, wait-free locking

ACM Reference Format:

Marvin Williams and Peter Sanders. 2025. The MultiQueue: A Simple and Fast Relaxed Concurrent Priority Queue. *ACM Trans. Parallel Comput.* 1, 1 (November 2025), 40 pages. <https://doi.org/10.1145/3771738>

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 882500).



Authors’ Contact Information: Marvin Williams, Karlsruhe Institute of Technology, Karlsruhe, Germany, williams@kit.edu; Peter Sanders, Karlsruhe Institute of Technology, Karlsruhe, Germany, sanders@kit.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

1 Introduction

The Priority Queue (PQ) is a fundamental¹ abstract data type that manages a dynamic collection of elements from an ordered universe. PQs support two basic operations: inserting an element and deleting the smallest element. They are central to many algorithms that benefit from dynamic reordering of operations, including job scheduling, graph searches, discrete event simulation, and best-first branch-and-bound searches.

Sequential PQs become a significant bottleneck in modern multi-core systems when used in parallel versions of these algorithms. PQs that permit the concurrent insertion and deletion of elements aim to address this issue but encounter semantic ambiguities not present in sequential PQs. These ambiguities arise when two threads attempt to delete the smallest element simultaneously, or when a deletion happens at the same time as the insertion of a new smallest element. *Linearizability* [20] is a strict consistency criterion for concurrent data structures that can resolve these ambiguities. Linearizability is a highly desirable property; several linearizable PQs have been proposed (e.g., [9, 24, 40]). Unfortunately, these PQs suffer from limited scalability due to inherent contention on the smallest element [14]. Moreover, parallel applications must handle non-optimal deletions, even when using linearizable PQs. To illustrate this, consider a parallel version of Dijkstra’s algorithm for shortest paths, where multiple threads concurrently delete nodes from the PQ. Unlike the sequential version, a node cannot be settled upon deletion from the PQ because another thread might simultaneously process a node yielding a shorter path to that node. This observation motivates the *relaxation* of the semantics of concurrent PQs, intentionally deleting non-optimal elements to improve throughput and scalability. The idea is that the increased throughput offsets the overhead of performing redundant or unnecessary work. Quality metrics quantify the degree of relaxation and can help assess the trade-off between quality and performance of a relaxed PQ.

In this article, we present the *MultiQueue*, a relaxed concurrent priority queue. We also evaluate it empirically and explore practical optimizations of its design. The MultiQueue distributes insertions and deletions across multiple internal (non-relaxed) priority queues to reduce contention. While the MultiQueue has a large design space, we focus on balancing scalability and quality. Specifically, our basic MultiQueue maintains a constant factor more queues than threads, inserts elements into random queues, and deletes elements from the better of two randomly chosen queues. To support arbitrary internal queues, these queues are protected by mutual exclusion locks. Despite the locks, the basic MultiQueue is probabilistically wait-free. It proves to be highly robust to different input distributions and operation sequences, but it has limited scalability due to poor data locality. We address this issue by using *buffers* for each internal queue and making threads *stick* to a subset of queues for multiple consecutive operations.

The article is organized as follows: In Section 2, we introduce notations and definitions used throughout this article. This includes the priority queue abstract data structure as well as the quality metrics *rank error* and *delay*. After discussing previous work on relaxed concurrent PQs and related topics in Section 3, we present the MultiQueue as the main contribution of this article in Section 4. In Section 5, we analyze the theoretical runtime and quality of the MultiQueue in terms of rank error and delay. We show that operations on the MultiQueue are nearly as fast as on sequential PQs, scaling linearly with the number of threads p . The rank errors and delays of the MultiQueue are linear in p in expectation and in $O(p \log p)$ with high probability. Additionally, we analyze branch-and-bound algorithms using relaxed PQs. In Section 6, we introduce various orthogonal enhancements to the MultiQueue to increase throughput. Section 7 addresses how to reliably detect termination for concurrent algorithms that finish when their PQ runs empty. In Section 8, we present an experimental evaluation of the MultiQueue, including a comprehensive comparison

¹It is covered in most introductory algorithm lectures and textbooks.

with alternative approaches found in the literature. We benchmark synthetic workloads to measure both maximum throughput and quality, as well as an adaptation of Dijkstra’s algorithm for shortest paths and a branch-and-bound algorithm for the knapsack problem. Section 9 concludes the article and discusses further research regarding the techniques used in the MultiQueue.

2 Preliminaries

Given a universe of elements \mathcal{U} and a strict weak ordering $<$ on \mathcal{U} , a *priority queue* (PQ) is a set of elements $Q \subseteq \mathcal{U}$ that supports the following operations:

insert Insert an element from \mathcal{U} into Q .

delete Remove and return a minimal² element from Q according to $<$.

The *rank* of an element $e \in \mathcal{U}$ with respect to a PQ Q is defined as $r(e) := |\{e' \in Q : e' < e\}| + 1$. For practical purposes, we also define the special element \perp that is greater than all elements in \mathcal{U} and cannot be inserted. It is returned by the delete operation in case the PQ is empty. *Concurrent* PQs allow multiple execution threads to call the insert and delete operations concurrently. Concurrent PQs introduce semantic ambiguities that do not arise in sequential PQs. For example, if multiple threads call delete simultaneously, it is unclear which elements should be removed. Similarly, if a delete operation occurs concurrently with an insert operation that inserts a new smallest element, it is unclear whether the new element should be considered for deletion. Linearizability [20] is a widely used consistency criterion for concurrent data structures in relation to sequential semantics. It requires that each operation appears to take effect instantaneously at some point between its invocation and response. However, this guarantee inevitably leads to high contention and limited scalability in linearizable PQ implementations [4]. To overcome this limitation, the semantics of concurrent PQs can be *relaxed*, allowing the delete operation to delete a non-minimal element or *fail* (i.e., return \perp without deleting an element) even when the PQ is not empty. Notably, this relaxation imposes no restrictions on when deletions are allowed to fail, making it impossible to determine whether the PQ is truly empty if a deletion fails. Most implementations of relaxed PQs therefore provide additional guarantees for the delete operation. For example, linearizable PQs are guaranteed to be empty if a delete operation, invoked after the completion of the last insert operation, fails.

A natural metric for the degree of relaxation is the *rank error*. For a deleted element e , the rank error is $r(e) - 1$, i.e., the number of elements in the PQ smaller than e . A complementary metric is the *delay*, measuring how many elements larger than e are deleted while e resides in the queue. For convenience, we define a failed deletion to have a rank error equal to the size of the PQ and a delay of zero. However, it still contributes towards the delay of all elements in the PQ. Note that a deletion with rank error r delays r elements by one. Thus, the sum of all delays is equal to the sum of all rank errors after all elements have been deleted. Nonetheless, the distribution of the delays can be an important performance indicator. For instance, if the minimum element is crucial for the progress of an algorithm, a PQ that consistently deletes the second-smallest element can lead to arbitrarily slow execution times. While the rank error will not exceed one, the delay of the minimum element can become arbitrarily large.

An algorithm is said to be *wait-free* if every thread is guaranteed to complete its operation in a bounded number of steps [19]. In the context of randomized algorithms, we say that an algorithm is *probabilistically wait-free* if the *expected* number of steps until a thread completes its operation is bounded. Unless otherwise stated, we denote p as the number of threads.

²Max-PQs that delete a maximal element work analogously.

An event occurs *with high probability* (w.h.p) if its probability of occurring approaches 1 as the number of threads p tends to infinity.

3 Related Work

This journal article has a long history. We introduced the basic idea to use the *two-choice paradigm* for concurrent PQs in the context of a preprint and SPAA brief announcement in 2014 [37, 38]. A “proper” conference paper [52] with considerable improvements like buffering and stickiness took until 2021. This journal article considers more stickiness variants, greatly expands the experiments (more benchmarks, machines, and competitors), discusses termination detection, and applies relaxed priority queues to parallel branch-and-bound algorithms. During this long time period, the set of authors changed.

There is now also some follow-up work building on previous publications on the MultiQueue. Alistarh et al. [1, 2] showed that the expected rank errors are in $O(p)$ and that the expected maximum rank errors are in $O(p \log p)$ if no elements smaller than a previously deleted element are inserted. Walzer and Williams [50] later offered a more precise generalization of these results. Postnikova et al. [34] proposed the *stealing MultiQueue* (SMQ) that has thread-local queues but occasionally steals a batch of elements from a random thread. This is conceptually similar to stickiness in the MultiQueue. The *Multi Bucket Queue* [55] enhances the efficiency of the MultiQueue design by grouping elements with the same priority. The Multi Bucket Queue additionally employs thread-local buffers to improve throughput.

There is a vast amount of work on sequential priority queues, focusing on different sets of operations and element types. Binary Heaps [51] are conceptually simple and have logarithmic worst-case running times in the number of elements, with favorable constant factors (at least when fitting in the cache). A natural generalization of binary heaps are k -ary heaps [22]—complete k -ary trees implicitly represented by an array in which elements are sorted along root-to-leaf paths. Heaps are available in the standard libraries of many programming languages, including C++ (`std::priority_queue`), Java (`java.util.PriorityQueue`), C# (`System.Collections.Generic.PriorityQueue`), Go (`container/heap`), Rust (`std::collections::BinaryHeap`), and Python (`heapq`). We also adopt heaps in our implementation and note that known PQs with better cache efficiency [33, 45] are problematic in our case, as they may occasionally require locking a PQ for a long time.

The oldest ancestor of the MultiQueue is the branch-and-bound algorithm by Karp and Zhang [23]. In their design, each processor owns a local PQ from which it deletes. Insertions go to the PQ of a random other processor. They show that, assuming synchronous execution, their algorithm expands at most $2e \approx 5.44$ times as many nodes as the sequential algorithm in expectation. However, this result does not transfer to the asynchronous setting, as shown by Alistarh et al. [2]. Our measurements confirm empirically that this PQ exhibits rank errors growing not only with p but also with the number of operations.

Sanders [42, 44] uses a similar design for a bulk-parallel priority queue. However, the globally best elements are determined using a fast parallel selection algorithm. This is asymptotically highly efficient but requires global synchronizations that are undesirable in many concurrent applications. The paper outlines an approach to make bulk-parallel PQs asynchronous by using an asynchronously managed buffer data structure. However, when this buffer runs empty, $\Theta(p)$ threads need to be activated to refill the buffer. This seems expensive on current architectures.³

³For example, in a concurrent hash table [25], reorganization is supported by having a pool of dedicated hardware threads. But tying up such a scarce resource considerably constrains the application. On the other hand, spawning new threads would incur prohibitive operating system overhead.

Hübschle-Schneider and Sanders [21] improve locality of bulk-parallel PQs by also inserting in the processor’s local PQ. By using a search tree for the local PQs, they can efficiently find the globally best elements in bulk-delete operations. Since the deleted elements may be distributed unevenly across the processors, explicit data redistribution is required.

Skip lists [35] have been intensively studied as a foundation for linearizable, lock-free concurrent priority queues (e.g., [8, 24, 47]). They organize elements into a hierarchy of sorted linked lists, where each list is a random sample of the list at the level below, and the lowest level contains all elements. This hierarchical structure accelerates search operations. However, linearizable PQs based on skip lists generally suffer from high contention, which limits their scalability. Rukundo and Tsigas [40] propose the *TSLQueue*, a lock-free linearizable PQ based on search trees and a sorted list. The TSLQueue empirically outperforms current skip-list-based linearizable PQs. However, the TSLQueue also suffers from bottlenecks and the worst-case performance is unclear, as its internal search tree can become unbalanced.

The *Spraylist* [3] is a relaxed concurrent PQ based on skip lists that mitigates the contention at the list head by *spraying* delete operations to elements close to the head. The spray is a biased random walk down the levels of the underlying skip list. Insertions can still cause heavy contention. The spraying yields rank errors in $O(p \log^3 p)$ with high probability. Wimmer et al. [53] propose the *k*-LSM, a relaxed concurrent PQ combining thread-local PQs and a global concurrent PQ, both implemented as log-structured merge-trees (LSMs) [33]. The local queues are bounded in size, and act as buffers for insertions and for elements “stolen” from other local queues. The analysis of the paper arrives at a rank error bound of $p \cdot k$ for buffers of size k [53]. However, it does not account for the possibly long merge-operations of the shared LSM, during which parts of it are inaccessible. The *contention avoiding priority queue* (CA-PQ) [41] by Sagonas and Winblad uses a global concurrent skip list and thread-local insertion and deletion buffers. The key idea is to dynamically switch between the global PQ and the local buffers depending on the contention. Under low contention, threads access the global PQ directly, and switch to the local buffers under high contention. This technique combines high accuracy in non-contended situations with high throughput under contention. Each thread accesses the global PQ after $m \in \Theta(p)$ operations on its local buffer to ensure that the quality does not deteriorate. The paper shows that a rank error in $O(p^2)$ is guaranteed for every m -th delete operation of each thread.⁴ Unfortunately, no guarantees have been proven for the remaining fraction of $1 - 1/m$ operations.

Another approach to low-overhead concurrent priority schedulers is grouping tasks into buckets such that tasks in the same bucket have the same priority. This approach allows for highly efficient data structures for managing the buckets, but is only applicable if the number of “active” priorities is small. One such application is Dijkstra’s algorithm with small integer edge weights. *Julienne* [11] is a framework for parallel graph algorithms where vertices are grouped into priority buckets. The buckets are processed in increasing order with a synchronization barrier between each bucket. With monotonically increasing elements, this is equivalent to sequential execution and thus work-efficient. The work-efficiency comes at the cost of poor parallelism and scalability if the buckets are small or some threads take much longer than others to reach the barrier. *OBIM* [32] and *PMOD* [54] maintain a global priority bucket map, but each thread works on its own version of the map asynchronously. In OBIM, threads update their buckets from the global map when they encounter an empty bucket. PMOD enhances OBIM by merging and splitting priority buckets dynamically, if they become too small or too large, respectively. Both OBIM and PMOD are highly scalable, but the quality of the processed elements can degrade over time [55].

Henzinger et al. [18] give a general framework for quantitative relaxation in the context of concurrent data structures. The rank error corresponds to the *out-of-order* relaxation in the framework, and the delay is related to the *lateness*

⁴Such a guarantee could also be achieved with a simplistic version of the MultiQueue in which each thread inserts and deletes from a fixed queue except that every m steps a thread scans all queues for the globally smallest element.

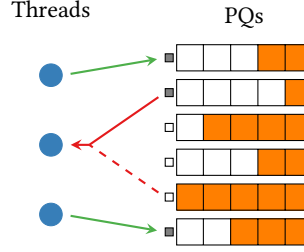


Fig. 1. Schematic view of the MultiQueue data structure with three threads, a queue factor of $c = 2$ and $d = 2$ deletion candidates. A gray square in front of a PQ indicates that the respective PQ is locked. Green and red arrows represent insertions and deletions, respectively. The dashed red line indicates the deletion candidate that is not deleted from.

relaxation. However, the lateness only considers the number of delete operations until a smallest element is deleted; delay is non-trivial to express within their framework. In a follow-up paper, they propose a data structure similar to the MultiQueue for a relaxed FIFO queue [17]. Rukundo et al. [39] introduce the more general concept of *semantic relaxation*. The idea is to build scalable, relaxed versions of sequential data structures by using multiple data structure instances, and threads reuse the same instance for a number of consecutive operations.

The two-choice paradigm is known to be a very powerful concept for load balancing without priorities [5, 6, 12, 31].

4 Basic MultiQueue

The MultiQueue comprises an array of sequential, non-relaxed PQs, each protected by a mutual exclusion lock⁵. The number of these *internal* PQs is proportional to the maximum number of threads p that may access the MultiQueue concurrently, resulting in $c \cdot p$ internal PQs for a *queue factor* $c > 1$. To insert an element, the MultiQueue selects an internal PQ randomly and attempts to lock it, repeating this process until locking succeeds. Then, it inserts the element into the locked PQ and releases the lock. Using the same strategy for deletions would cause rank errors to degrade not only with p but also with the number of operations (see Figure 7 and [2]). To counteract this, the delete operation requires more effort: instead of selecting a single PQ, the MultiQueue selects $d \geq 2$ PQs randomly and attempts to lock the one containing the smallest element among them.

Figure 1 depicts the MultiQueue structure, and Figure 2 gives pseudocode for the insert and delete operations. Note that the operations never wait for a lock to be released. There are always unlocked PQs available, since at most p PQs are locked at any given time. We emphasize that, for deletions, only the “better” PQ is locked, which requires that reading the smallest elements of a PQ is thread-safe. For implementation details, refer to Section 8.1. The delete operation returns \perp if both selected PQs are empty, optimistically reporting that the MultiQueue is likely empty without considering additional PQs. In particular, as with most relaxed priority queues (e.g., k -LSM, SMQ, CA-PQ) failed deletions are not linearizable. In Section 7, we describe how to adapt the delete operation such that applications can reliably detect when the MultiQueue is empty with low overhead.

The motivation for considering multiple PQs for deletion stems from previous work on randomized resource allocation, particularly the balls-into-bins process. In this process, n balls are successively placed into one of p bins. The number of balls in a bin is the *load* of this bin, and the normalized load is defined as the load minus the average load across all bins. Assuming $n \gg p$, the normalized maximum load at the end of this process is in $\Theta(\sqrt{n \log p/p})$ with high

⁵One can also use thread-safe (linearizable) PQs to avoid locking.

<p>Procedure: insert</p> <p>Data: Element to insert e</p> <pre> repeat $I \leftarrow \text{random}(0, c \cdot p - 1)$ until tryLock($A[I]$) $A[I].\text{insert}(e)$ unlock($A[I]$) </pre>	<p>Procedure: delete</p> <p>Result: Removed element or \perp</p> <pre> repeat $I_1 \leftarrow \text{random}(0, c \cdot p - 1)$ $I_2 \leftarrow \text{random}(0, c \cdot p - 1)$ if $A[I_2].\text{min}() < A[I_1].\text{min}()$ then $I_1 \leftarrow I_2$ until tryLock($A[I_1]$) $e \leftarrow A[I_1].\text{delete}()$ unlock($A[I_1]$) return e </pre>
--	---

Fig. 2. Pseudocode for the insert and delete operations with $d = 2$. The internal PQs are stored in array A . The min function returns a smallest element or \perp if the PQ is empty.

probability, provided that each ball is assigned to a bin uniformly at random [36]. If, instead, the least-loaded of $d \geq 2$ randomly chosen bins is selected for each ball, the normalized maximum load is in $O(\log \log p)$ with high probability [6], yielding an exponential improvement known as the *two-choice paradigm* [31]. Alistarh et al. [2] propose a protocol in which, with probability $0 < \beta \leq 1$, two PQs are considered for deletion, and only one otherwise, in order to achieve a smaller effective value of $d = 1 + \beta$.

In the time window during deletions, between comparing the smallest elements of the selected PQs to determine which PQ to lock and successfully locking it, the comparison can become *stale*, meaning that the smallest element in the selected PQs has changed. A comparison becomes stale if another thread deletes the smallest element from one of the candidate PQs or inserts a new smallest element into one of them during this time window. Due to stale comparisons, the wrong PQ may be locked, resulting in a “bad” deletion or even a failed deletion, despite the other selected PQs being nonempty.⁶ To avoid stale comparisons, one can lock both PQs before comparing them. This approach introduces significant overhead and requires a queue factor of $c > 2$ to maintain wait-freedom. Alternatively, one can re-compare the PQs after acquiring the lock and retry the operation if the comparison yields a different outcome than before. Preliminary experiments showed no significant quality improvements from preventing stale comparisons, suggesting that they are rare or have a negligible impact in practice. Alistarh et al. confirm, for a simplified MultiQueue process, that the quality of deleted elements does not deteriorate with occasional stale comparisons [1].

5 Theoretical Analysis

In this section, we analyze the theoretical performance and quality of the MultiQueue. We also analyze the running time of branch-and-bound algorithms when using relaxed concurrent PQs instead of sequential PQs.

5.1 Running Time

We analyze the asymptotic running time of the insert and delete operations within a realistic asynchronous shared memory model (specifically, the aCRQW model [46, Section 2.4.1]). In this model, the time required for performing p contended writes to the same machine word is in $O(p)$.

⁶This issue does not occur when comparing and locking the PQs atomically, e.g., using transactional memory.

THEOREM 5.1. *The expected time for a thread to acquire a lock during the insert and delete operations is in $O(1)$.*

PROOF. At most $p - 1$ internal PQs are locked by other threads at any given time. Hence, the worst-case probability r that a randomly selected PQ is locked is

$$r := \frac{p-1}{cp} \leq \frac{1}{c} < 1.$$

Since the probabilities of subsequent locking attempts to succeed are independent, the probability that a thread needs i attempts to find an unlocked queue is $r^{i-1}(1-r)$. Let X be the random variable indicating the number of attempts to find an unlocked queue. The expected number of attempts to find an unlocked queue is therefore

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \cdot r^{i-1}(1-r) = \frac{1}{1-r} \leq \frac{1}{1-\frac{1}{c}} = \frac{c}{c-1} \in O(1).$$

Since the expected number of threads contending to lock a PQ is constant, the same bound holds for the expected execution time. \square

We obtain the following bounds for the comparison-based model and for integer keys:

COROLLARY 5.2. *A MultiQueue with binary heaps needs constant average insertion time and expected time $O(\log n)$ per operation for worst-case operation sequences. With van Emde Boas trees [29, 49] and integer keys in $\{0, \dots, U-1\}$, an expected time of $O(\log \log U)$ per operation is achieved.*

Theorem 5.1 holds even when other threads are suspended, since in the worst case, each suspended thread holds one lock. The overhead of attempting to acquire these locks is already accounted for in the analysis. As a result, each thread is expected to make progress in a bounded number of steps, making the MultiQueue probabilistically wait-free. Note that “progress” in this context refers to completing an insert or a (potentially failing) delete operation, rather than the completion of an application using the MultiQueue. Specifically, if a thread is suspended while holding a lock on an internal PQ, the elements in that PQ become inaccessible to other threads. This can potentially delay an application’s progress indefinitely, rendering it not wait-free.

5.2 Quality

A general quality analysis of the MultiQueue is still an open problem. Alistarh et al. were the first to provide an asymptotic analysis of the rank error in a simplified sequential MultiQueue process, where no elements smaller than any previously deleted element are inserted. They show that the expected rank error is in $O(p)$ and the expected maximum rank error is in $O(p \log p)$ for any number of deletions. They later showed that these bounds hold even in a concurrent setting with stale comparisons [1, 2]. Walzer and Williams [50] present an exact analysis of the rank error distribution for the same sequential process with more general deletion strategies. The expected long-term rank error for $d = 2$ turns out to be $\frac{5}{6}cp - 1 + \frac{1}{6cp}$. Both Alistarh et al. [2] and Walzer and Williams [50] confirm that the distribution of elements to the PQs does not stabilize for $d = 1$, leading to diverging rank errors (see Figure 7).

In this section, we will explain how *rank errors* and *delays* can be estimated under simplified but intuitive assumptions. In the following, we only consider the sequential execution of operations. We assume that, at any given time, each element is equally likely to be in any queue. This assumption holds at least until the first delete operation is executed. The reasoning behind this assumption is as follows: Insertions place elements into random queues, which moves the system toward a uniform distribution of elements. Deletions, on the other hand, select PQs with small elements more frequently than those with large elements, thus controlling the deviation among the queues. Walzer and Williams [50]

show that this assumption does not generally hold but conjecture that the distribution of elements will never be “too far off” the uniform distribution, even for arbitrary operation sequences.

Rank error. Let R be the random variable indicating the rank error of a delete operation. There are R elements smaller than the deleted element, and all of them are in PQs other than the ones considered for the delete operation. Under our assumptions, the probability for a specific element to be in one of the selected PQs is $z := \frac{d}{cp}$, independent of the other elements. Thus, R follows a geometric distribution with parameter z :

$$P(R = i) = (1 - z)^i z.$$

Since $E[R] = \frac{1}{z} - 1$, the expected rank error is in $O(p)$.

The probability to delete an element with rank error i or greater is given by

$$P(R \geq i) = (1 - z)^i.$$

For a rank error $l = \frac{1}{z} \log p \in O(p \log p)$, we have $P(R \geq l) = (1 - z)^l \approx e^{-z \cdot l} = p^{-1}$. Therefore, rank errors are in $O(p \log p)$ with high probability.

Delay. Since deleting an element with rank error i delays i elements, the sum of all delays is equal to the sum of all rank errors after all elements are deleted. Therefore, the expected delay is equal to the expected rank error. Consider an element e with rank $i + 1$. The probability for a deletion to remove an element larger than e , thus increasing the delay of e by one, is $P(R > i)$. With probability $P(R = i)$, the delete operation removes e . Hence, the probability that a delete operation either increases the delay of e or removes e is $P(R \geq i)$. The conditional probability that a delete operation, which would otherwise delay e , instead removes e is

$$P(R = i \mid R \geq i) = \frac{P(R = i)}{P(R \geq i)} = \frac{(1 - z)^i z}{(1 - z)^i} = z.$$

Since this probability is independent of i , the delay also follows a geometric distribution. In fact, the delay distribution is identical to the rank error distribution. Consequently, the expected delay is in $O(p)$ and, with high probability, the delay is in $O(p \log p)$.

5.3 Branch-and-bound

In this section, we analyze the run-time of the best-first branch-and-bound scheme using a relaxed concurrent PQ. Our approach is similar to the analysis for bulk-parallel priority queues [42], but leverages the expected delay to estimate the overhead introduced by the relaxation.

Branch-and-bound algorithms solve optimization problems by exploring and pruning the search tree of possible (partial) solutions. Given the currently best solution as an upper bound for the optimal (minimal) solution, partial solutions with lower bounds larger than the currently best solution are pruned. The best-first heuristic selects the partial solution with the smallest lower bound for further exploration. This heuristic is typically implemented by a PQ managing partial solutions ordered by their lower bound.

Let H be the search tree of partial solutions, and let $H_{\leq} \subseteq H$ be the subset of partial solutions whose lower bounds are not larger than the optimal solution o . A sequential best-first branch-and-bound algorithm might explore all of H_{\leq} while pruning all partial solutions in $H_{>} := H \setminus H_{\leq}$. Partial solutions in $H_{>}$ cannot be explored before o is found, since their lower bounds are larger than o . After o is found, all partial solutions in $H_{>}$ are pruned. However, with relaxed PQs, some nodes from $H_{>}$ might be explored before they can be pruned. Figure 3 illustrates the search tree and the

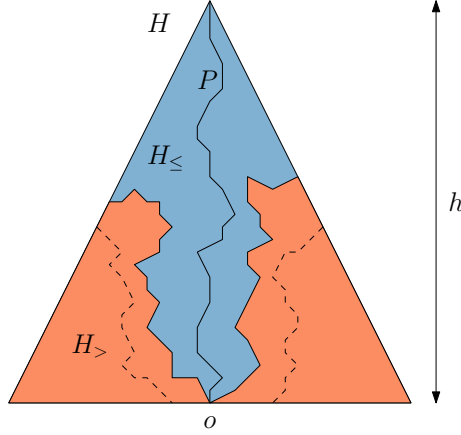


Fig. 3. Schematic view of the search tree H with height h with nodes in H_{\leq} and $H_{>}$. The path P from the root to the optimal solution o is highlighted. The dashed lines bound the area of $H_{>}$ that are explored due to the delay.

nodes that are unnecessarily explored due to the delay. Let P be the path from the root of H to o , h be the length of P , and D the expected delay of the relaxed PQ. Until o is found, the PQ always contains at least one partial solution from P . Each deletion from $H_{>}$ delays all partial solution on P that are in the PQ at the time of the deletion. Since each partial solution in P is delayed by D in expectation, at most Dh nodes from $H_{>}$ are expected to be explored before o is found. In total, $n' \leq H_{\leq} + Dh$ nodes are explored in expectation. Assuming that PQ operations and processing a node take $O(\log n)$ time, the expected parallel run-time T_{par} with p threads is

$$T_{\text{par}} = \frac{T_{\text{seq}}}{p} + O\left(\left(\frac{H_{\leq} + Dh}{p}\right) \log n\right).$$

For the MultiQueue, D is in $O(p)$, leading to a total expected run-time of

$$T_{\text{par}} = \frac{T_{\text{seq}}}{p} + O\left(\left(\frac{H_{\leq}}{p} + h\right) \log n\right).$$

Note that this bound is essentially the same as for bulk-parallel PQs [42]. However, a crucial difference is that with bulk-parallel PQs, the bound applied only when node processing always took the same time.

6 Practical Improvements

In this chapter, we introduce practical improvements to the MultiQueue design as presented in Section 4. Even for small numbers of elements, cache efficiency is an issue for the MultiQueue. By operating on a specific internal PQ, a thread will move the corresponding data into its cache. Most likely, this PQ will be accessed by other threads next, leading to cache misses for these threads as well as invalidation traffic due to cache coherency. We address this issue and present techniques to significantly improve the cache efficiency of the MultiQueue in the following sections.

6.1 Buffering

To reduce the average number of cache lines accessed by the insert and delete operations, we enhance each internal PQ with an *insertion buffer* and a *deletion buffer* of fixed capacity. Generally, elements are inserted into the insertion buffer and deleted from the deletion buffer. We maintain the invariants that the smallest elements are in the deletion

<p>Procedure: insertWithBuffers</p> <p>Data: Element to insert e</p> <pre> if $I = \emptyset \wedge Q = \emptyset \wedge D < C_D$ then $D \leftarrow D \cup \{e\}$ return if $e < \max(D)$ then if $D < C_D$ then $D \leftarrow D \cup \{e\}$ return $e' \leftarrow \max(D)$ $D \leftarrow (D \setminus \{e'\}) \cup \{e\}$ $e \leftarrow e'$ if $I = C_I$ then flush I into Q $I \leftarrow I \cup \{e\}$ </pre>	<p>Procedure: deleteWithBuffers</p> <p>Result: Removed element or \perp</p> <pre> if $D = \emptyset$ then return \perp $e \leftarrow \min(D)$ $D \leftarrow D \setminus \{e\}$ if $D = \emptyset$ then flush I into Q refill D from Q return e </pre>
---	---

Fig. 4. Pseudocode for inserting into and deleting from a locked PQ Q with insertion buffer I and deletion buffer D . The buffers have capacities C_I and C_D , respectively. The \max and \min operations return a largest and smallest element or \perp if the set is empty, respectively. Refilling D from Q is done by iteratively deleting the smallest element from Q and inserting it into D until D is full or Q is empty.

buffer in sorted order, and if the deletion buffer is empty when trying to delete an element, the insertion buffer and the PQ are also empty. Therefore, small elements have to be inserted into the deletion buffer directly. If the insertion buffer is full, its elements are flushed into the PQ. If the deletion buffer becomes empty, it is refilled with the smallest elements from both the insertion buffer and the PQ. This is done by first flushing the insertion buffer into the PQ and then refilling the deletion buffer solely from the PQ. Figure 4 shows the pseudocode for the `insertWithBuffers` and `deleteWithBuffers` operations. For implementation details of the buffers, refer to Section 8.1. An alternative approach to refill the deletion buffer is to refill from the PQ first, and then swap elements with the insertion buffer to re-establish the invariant that the smallest elements are in the deletion buffer. This approach has the advantage that all interactions between the buffers and the PQ are in batches of fixed size, which can be exploited by specialized PQs (see Section 6.2). However, it incurs additional moves of elements between the buffers and the PQ.

With buffers, insertions and deletions interact with the internal PQ itself only when the buffers are full or empty. An insertion typically reads the largest element from the deletion buffer to determine whether the element should be inserted into the deletion buffer directly, and otherwise operates only on the insertion buffer. Deletions typically only access the deletion buffer. Buffering can reduce cache misses, since only one thread accesses the PQ in a bulk-fashion, exhibiting high temporal locality without interference from other threads. Implicit heaps, for example, exhibit high locality for insertions, and at least exhibit some locality near the root and at the rightmost end of the bottom layer of the tree for deletions. Further, buffering retains the semantics of the internal PQs, and thus does not affect the rank error or delay in theory. However, buffering increases the worst-case time for each operation. If an operation locks a PQ that contains small elements for a long time, the delay of those elements, as well as the rank error of other deletions, increases.

6.2 Cache-efficient PQs

Implicit tree-like data structures like binary heaps can be used to avoid cache misses from indirection within the internal PQs. k -ary heaps with $k > 2$ improve on binary heaps by reducing the number of cache misses to $O(\log_k n)$ if k

elements fit into one cache line. To fully exploit the fact that the MultiQueue accesses the internal PQ in a bulk-fashion due to buffering, data structures that directly support batch operations can also be utilized. A promising data structure is the *merging binary heap*, an adaptation of the *parallel heap* by Deo and Prasad [10]. Merging binary heaps are structured like binary heaps, but each node contains a fixed number of sorted elements. The heap invariant is that the first element of each node is not smaller than the last element of its parent node. Insertion and deletion work similarly to ordinary binary heaps.

New nodes are inserted as a leaf. Then, the node is merged with its parent, moving the smaller half into the parent node and the larger half into the child node. This process is repeated with the parent of the current node until the first element of the node is not smaller than the last element of its parent or the root is reached. When the root node is deleted, the two child nodes are merged and the smaller half is put into the parent node, the larger half into the child node which had the larger last element. Then, the other child node is empty, and the process is repeated with this child until a leaf is reached. Finally, the last node is moved to this leaf, and the heap invariant is restored analogously to the insertion process. While merging binary heaps require fewer tree operations than k -ary heaps, they come with additional algorithmic complexity and higher worst-case access times. In our preliminary experiments, merging binary heaps and k -ary heaps performed very similarly, so we decided to use the conceptually simpler k -ary heaps.

6.3 Stickiness

Stickiness is a simple yet effective technique to increase temporal cache locality at the cost of potentially higher rank errors and delays. The basic idea is as follows: Threads stick to the same set of $d \geq 2$ internal PQs for $s \geq 1$ consecutive operations (the *stickiness period*). Insertions pick one PQ from the set at random. Deletions use the PQ with the smallest element among the set. PQs are still only locked for single operations. Each thread replaces its set with a new set of d PQs after s consecutive operations or if locking fails for one operation (to avoid blocking). We assume $d \leq c$ to ensure that there are always enough PQs for each thread to stick to.

In the *simple* variant, the PQs are chosen uniformly at random, independently for each thread. Thus, multiple threads may stick to the same PQs simultaneously, which is undesirable because of contention and cache invalidations. To mitigate this issue, threads can leave a thread-specific mark in the lock of their most recently accessed PQ. Then, they can detect if they try to lock a PQ that another thread recently accessed, and choose a different one to stick to.

The *swap* variant helps guiding threads to currently unused PQs by maintaining a permutation of PQ indices in a global array. Each thread is assigned d fixed positions in this array, indicating the PQs to stick to. Whenever a thread wishes to replace its set of PQs, it atomically swaps the indices in each of its array positions with random other positions. Thus, no two threads stick to the same PQ simultaneously. Still, threads need to acquire locks for each operation because there might be ongoing operations from other threads on the assigned PQs. We use the algorithm given in Figure 5 to atomically swap entries in the permutation array, since most hardware does not natively support swapping values between two memory locations atomically. It is crucial for the correctness of the algorithm that each entry can only be invalidated (set to -1) by one specific thread. The algorithm is probabilistically wait-free because valid entries are found in expected constant time. Our preliminary experiments showed that the marking approach is inferior to the *swap* variant.

Another approach we did not investigate further is to induce the assignment using a global permutation of the form $\pi(i) = i \cdot a + b \bmod m$ with parameters $a \in \mathbb{N}$ and $b \in \mathbb{N}_0$ such that a and $m = c \cdot p$ are co-prime.⁷ The parameters a and

⁷Since a and m are co-prime, there exists a unique a^{-1} such that $aa^{-1} \equiv 1 \pmod m$. For $\pi(i) = \pi(i')$ it must be $ia + b \equiv i'a + b \pmod m \Rightarrow ia \equiv i'a \pmod m \Rightarrow iaa^{-1} \equiv i'aa^{-1} \pmod m \Rightarrow i \equiv i' \pmod m$.

```

Procedure: atomicSwap
Data: Index to swap from  $i$ 
 $a_i \leftarrow \text{atomicExchange}(A[i], -1)$ 
repeat
   $j \leftarrow \text{random}(0, c \cdot p - 1)$ 
   $a_j \leftarrow A[j]$ 
until  $a_j \neq -1 \wedge \text{compareAndSwap}(A[j], a_j, a_i)$ 
 $A[i] \leftarrow a_j$ 

```

Fig. 5. Pseudocode for the swapping an entry i with another randomly chosen entry in the permutation array A . The operation $\text{atomicExchange}(A, v)$ atomically reads the value at A and sets it to v , the operation $\text{compareAndSwap}(A, e, x)$ atomically compares the value at A with e and sets it to x if they are equal.

b are shared between all threads. Each thread is assigned fixed unique indices $i, \dots, i + d - 1 \in [0, m - 1]$, and sticks to the PQs $\pi(i), \dots, \pi(i + d - 1)$. This approach avoids the rather expensive swapping in the permutation array, but it changes the PQ assignments for all threads simultaneously. The threads have to agree when to change the permutation, which can be challenging in practice.

7 Termination Detection

Termination detection is the process of determining when an algorithm has completed its task. Many algorithms involving a priority queue repeatedly delete and process (thereby possibly inserting new elements) elements from the PQ until the PQ becomes empty. Dijkstra's algorithm is a well-known example for this pattern. While it is trivial to detect the termination condition in sequential settings, it can be challenging in parallel settings.⁸ The goal is to ensure that all threads keep working until the PQ is permanently empty and then terminate. Linearizable failed deletions or an `isEmpty` operation generally do not suffice, because a thread may observe an empty PQ while other threads are still processing elements. Thus, proper termination detection requires some kind of coordination between the threads. If the total number of elements to process is known in advance, each thread can maintain a local counter of processed elements. When a thread fails to delete an element, it increments a global atomic counter by the number of elements it has processed since it last updated the counter. Then, threads can terminate as soon as the counter reaches the total number of elements to process. However, if the total number of elements to process is unknown, more sophisticated coordination is required. Tracking the number of remaining elements to process with a global atomic counter is infeasible in high-throughput scenarios due to the high contention on that counter. Maier et al. [25] propose a more scalable version of this approach in the context of approximating the size of concurrent hash tables, where each thread counts the number of its insertions and deletions, and update a global counter sporadically. However, this counter can become zero despite some elements remaining to be processed, making it unsuitable for termination detection. Ellen et al. [15] propose a scalable concurrent non-zero indicator for the number of remaining elements to process, incurring a small overhead per operation. This indicator is especially useful in scenarios where the PQ becomes empty frequently. However, relaxed concurrent PQs are most useful in scenarios where the PQ contains many elements for most of the time and processing elements is very fast, making this approach less suitable. If there are few elements to process most of the time, faster unordered data structures might be more suitable than relaxed PQ.

⁸Termination (or *quiescence*) detection is a well-studied problem in distributed systems. For an overview, see [27] and [28].

Procedure: ProcessUntilEmpty
Data: Priority queue Q
Data: Number of threads p
Data: Global atomic counters polling and idle

```

1 repeat
2    $e \leftarrow Q.delete()$ 
3   if  $e = \perp$  then
4     polling  $\leftarrow$  polling + 1
5     repeat
6        $e \leftarrow Q.delete()$ 
7       if  $e \neq \perp$  then break
8       if polling <  $p$  then continue
9       idle  $\leftarrow$  idle + 1
10      while polling =  $p$  do
11        if idle =  $p$  then return
12      end
13      idle  $\leftarrow$  idle - 1
14    end
15    polling  $\leftarrow$  polling - 1
16  end
17  // process  $e$ 
18 end

```

Fig. 6. Pseudocode for termination detection. polling and idle are global atomic counters, initialized to 0.

Another approach is to periodically synchronize all threads and determine whether all elements have been processed. Since global synchronization is expensive, this approach must be carefully designed to avoid excessive synchronization while ensuring that all threads terminate in a timely manner.

The algorithm given in Figure 6 avoids global synchronization and incurs no overhead for successful deletions, relying on the following assumption: The PQ is guaranteed to be empty if the last operation of every thread was a failed deletion and currently no insertions are in progress. To the best of our knowledge, most relaxed PQ designs can be implemented such that this assumption holds. Note that linearizable PQs provide an even stronger guarantee by definition: The PQ is empty if for any thread a delete operation failed that started after the last insert operation completed. However, the delete operation of the MultiQueue as given in Figure 2 does not satisfy this assumption. We therefore extend it to retry the operation a fixed number of times upon failure, after which all internal PQs are scanned linearly for elements (locked PQs are skipped).

The algorithm detects termination by tracking the number of threads whose last operation was a failed deletion. To prevent premature termination, the algorithm counts the threads without work twice, similar to the well-known *double counting* mechanism for distributed systems [28]. Threads can be in one of three states: *working*, *polling*, or *idling*. The number of threads in the polling and idling states are tracked using dedicated counters. All threads begin in the working state, where they repeatedly delete elements from the PQ (Line 2). If a thread fails to delete an element, it transitions to the polling state (Line 4). While polling, the thread keeps attempting deletions until either it succeeds, or all other threads are also polling (Lines 6–8). If a deletion was successful, the thread returns to the working state. If a polling thread detects that all threads are polling, it transitions to the idling state (Line 9). Note that terminating at this

point would be premature, since another polling thread may have just successfully deleted an element but has not yet entered to the working state. In the idling state, a thread waits until all other threads also become idle, at which point it terminates (Line 11). However, if another thread transitions back to the working state, the idling thread re-enters the polling state.

In the following, we argue that the termination detection algorithm keeps all threads working until the PQ is empty and no new elements will be inserted, and that all threads eventually terminate under these conditions. A thread terminates if and only if it observes the idle counter reaching p in Line 11. After a thread executed Line 9 (increment the idle counter), it must execute Line 12 (decrement the idle counter) before it can execute Line 9 again. Thus, for the idle counter to reach p , all threads must be in the idle state at the same time. At that time, every thread must have performed a failed deletion without subsequently starting an insertion. Given the assumption that the PQ is empty under these conditions, and no thread has an element to process, all threads can safely terminate. Moreover, each thread attempts to delete the next element to process within a bounded number of steps (assuming uniform progress and instantaneous memory propagation) until it terminates. If no elements remain to be processed, all threads will eventually leave the working state and never re-enter it. Thus, the polling counter will eventually reach p , and all threads will eventually transition to the idle state and never leave it again. Once all threads reach the idle state, the idle counter reaches p and all threads terminate.

The algorithm is not lock-free, as all threads must increment the polling and idle counters before any thread can terminate. If a thread halts before completing both increments, it delays the termination of all other threads. However, the algorithm does allow threads to leave and join the system at specific points. Note that arbitrarily suspending threads would break most applications, regardless of the termination detection algorithm. For example, a thread could be suspended right after deleting an element but before processing it. To avoid this issue, deleting and processing an element (including possible insertions) would need to occur within a single atomic transaction, which is beyond the scope of this work.

8 Experiments

The experiments are divided into three parts. In the first part, we measure the rank errors and delays exhibited by the basic MultiQueue and compare them to the results of the theoretical analysis. In the second part, we perform parameter tuning of the MultiQueue, including the practical improvements, to obtain a set of configurations that offer a good trade-off between throughput and quality. Finally, in the last part, we conduct an extensive comparison of the MultiQueue with its competitors, evaluating performance across various workloads and different hardware architectures. This includes stress tests and concurrent variants of Dijkstra’s algorithm and a branch-and-bound algorithm. The latter might be of independent interest as parallel algorithms in their own right.

8.1 Implementation

At its core, the MultiQueue is an array with one entry for each internal PQ. Each entry contains a mutual exclusion lock, the insertion and deletion buffers, a redundant copy of the smallest element, and a pointer to the PQ itself. Thus, the array requires a fixed amount of memory in $O(p)$. Comparing the smallest elements of internal PQs is done by comparing the redundant copies, which does not require locking the PQs. Entries and the buffers themselves are padded and aligned to *cache lines* to prevent false sharing of neighboring entries. Thread-specific data, such as stickiness counters, are stored in thread-local storage.

The insertion buffer consists of an array and a size counter. The deletion buffer is implemented as a ring buffer with elements sorted in ascending order. Ring buffers support removing the first element in constant time, while insertions at arbitrary positions take at most $n/2$ element moves with n elements. The PQs themselves are implicit k -ary heaps stored in dynamically growing arrays.

Our implementation of the MultiQueue data structure allows the user to easily exchange the PQ implementation and the memory allocation strategy. The priority queue implementation we use for our experiments is built on top of `std::vector` using the default `std::allocator` from the C++ standard library. Memory is only returned to the system when the PQ is destructed. Thus, the overall memory consumption is in $O(p + n)$, where n is the maximum previous size of the queue.

We use the extended `delete` operation that retries a fixed number of times before falling back to a linear scan of the internal PQs (see Section 7). While our implementation can handle generic element types, it is designed and optimized for elements of small size that are cheap to copy and compare.

8.2 Methodology

Stress Tests. The stress tests are designed to measure the maximum throughput and quality under high contention. For compatibility with all tested priority queue implementations, elements are key-value pairs of two unsigned 64-bit integers. The order of the elements is determined by the key. The value of each element is a globally unique identifier to ensure that no two elements are equal. The following stress tests are conducted:

Monotonic: At first, n elements with keys $\{1, \dots, n\}$ are inserted into the PQ to simulate a busy system (the *pre-fill*).

Then, each thread repeatedly performs alternating `delete` and `insert` operations until a fixed number of total iterations is reached. The keys of inserted elements are drawn uniformly at random from the range $[k, k + n]$, where k is the key of the previously deleted element.

Insert-Delete: For some number of elements n , all threads insert elements with keys drawn uniformly at random from the range $[1, n]$ until n elements are inserted. Then, all threads perform `delete` operations until the PQ is empty.

The monotonic stress test keeps the total number of elements in the PQ constant and roughly mimics applications where the elements monotonically become larger, such as Dijkstra’s algorithm and branch-and-bound algorithms. The rank of newly inserted elements is more evenly distributed compared to the naive approach to draw keys uniformly at random from the range $[1, n]$, where the ranks of newly inserted elements are very low most of the time. We consider the monotonic stress test as reasonably general to assess the general throughput, rank errors and delays of concurrent PQ implementations. The insert-delete stress test complements the monotonic test as it measures the maximum throughput of insertions and deletions separately with varying numbers of elements. The insert-delete workload is reminiscent of heap-sort.

Measuring Rank Errors and Delays. Measuring the rank errors and delays in real-time imposes the practical problem that we need to know which elements are present in the PQ at the time of each `delete` operation. We approach this problem as follows. Each thread logs its operations locally with timestamps using a low-overhead high-resolution clock.⁹ Insertion-timestamps are recorded immediately *before* each `insert` operation and deletion-timestamps are recorded immediately *after* each `delete` operation. To compute the rank error and delays, we merge the logs of all threads to

⁹The resolution of the `std::chrono::high_resolution_clock` from the C++ standard library was sufficient on our platforms. The timestamp counter (TSC) might have higher resolution but is not guaranteed to be synchronized across cores and/or sockets.

one global sequence of operations sorted by timestamp. We then replay this sequence sequentially and compute the exhibited rank errors and delays for each delete operation. While this approach is not perfectly accurate, we deem it sufficient for our purposes. Notably, even linearizable PQs can exhibit rank errors due to the time window between the timestamps and the actual linearization points. We use an augmented B+ tree (based on the `tlx::btree_map` from the `tlx` library [7]) to replay the operations and compute the rank errors and delays. The B+ tree holds the elements in priority order and contains exactly the elements that were in the priority queue at the point in time of the currently replayed operation according to the timestamps. By augmenting the inner nodes of the B+ tree with their size, the rank of deleted elements can be determined in logarithmic time (e.g., [46, Section 7.5]). Each node v is further augmented with a delay counter d_v . The delay of an element e (leafs in the B+ tree) is the sum of the delay counters on the path from the root to e . Since the delay of newly inserted elements is zero, d_e is initially set to the negated sum of all delays on the path from e to the root. To increase the delay of smaller elements when deleting an element, the delay counters of nodes as close to the root as possible are incremented. When performing balancing operations on the tree, the delay counters of the manipulated nodes are propagated downward to unchanged subtrees.

Setup. The MultiQueue and all benchmarks are implemented in C++17. Competitors are implemented in C and C++. All code is compiled with GCC 12.1 using `-O3 -DNDEBUG -march=native`. We use three different machines for the experimental evaluation:

Machine AMD is equipped with an AMD EPYC 7702 processor (x86-64 ISA) and 1 TB of DDR4 RAM. The processor hosts 64 cores with 2 hardware threads each. The maximum clock frequency is 3.35 GHz. Each core has 32 kB of L1 data cache and 1 MB of L2 cache. Four cores share 16 MB of L3 cache. The machine runs Ubuntu Server 22.04 with Linux 5.4.

Machine Intel-NUMA is equipped with four Intel Xeon Gold 6138 processors with 188 GB of DDR4 RAM each. Each processor has 20 cores with 2 hardware threads each. The maximum clock frequency is 3.7 GHz. Each core has 32 kB of L1 data cache and 1 MB of L2 cache. Each NUMA node shares 27.5 MB of L3 cache. The machine runs Ubuntu Server 22.04 with Linux 5.4.

Machine ARM is equipped with a Neoverse-N1 processor (AArch64 ISA) and 256 GB of DDR4 RAM. The processor hosts 80 cores. The maximum clock frequency is 3 GHz. Each core has 64 kB of L1 data cache and 1 MB of L2 cache. The machine runs Ubuntu Server 22.04 with the Linux 5.15 kernel.

We conduct the parameter tuning and all comparison experiments on machine AMD, as it is compatible with all implementations and offers the most homogeneous scalability behaviour. Additionally, the stress tests are performed on the Intel-NUMA and ARM machines. The throughput of the monotonic stress test is measured over 2^{24} iterations per thread, or until a timeout of 5 s is reached. Whenever possible, sufficient memory is pre-allocated to prevent memory allocation during the benchmarks. Each stress test is repeated 5 times and the mean is reported. Execution threads are pinned to distinct hardware threads, the clock frequency is fixed and frequency boosting is disabled. We use the `numactl` tool on the Intel-NUMA machine to distribute the data evenly across the participating NUMA nodes. Cache misses are measured on the AMD machine programmatically with the *L2 Cache Miss from Data Cache Miss* performance counter¹⁰ using the PAPI¹¹ library.

¹⁰<https://www.amd.com/content/dam/amd/en/documents/processor-tech-docs/programmer-references/55803-ppr-family-17h-model-31h-b0-processors.pdf>

¹¹<https://icl.utk.edu/papi/>

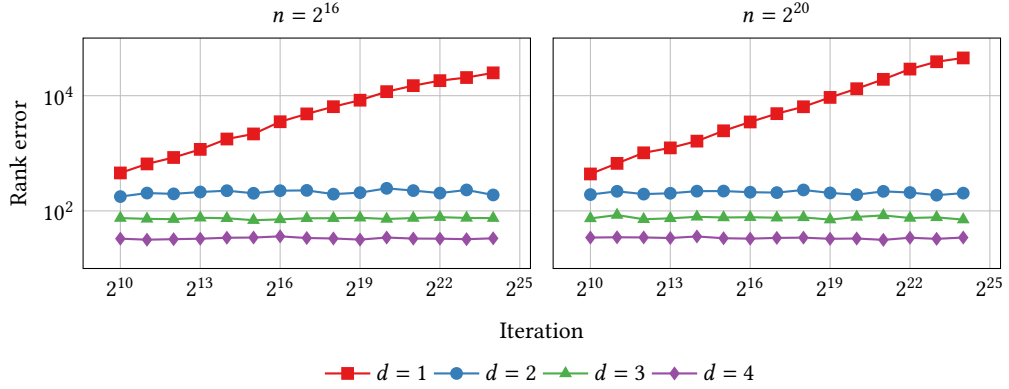


Fig. 7. Development of the rank error during the monotonic stress test for 256 PQs with different pre-fills n . Each line shows a different number of candidate PQs d , and each data point shows the mean of the previous 1024 iterations.

8.3 Quality of the basic MultiQueue

In this section we want to measure the quality characteristics of the MultiQueue design itself and compare it to the results from the theoretical analysis. In the following, we refer to the analysis given in Section 5.2 as the estimation and the results from Walzer and Williams [50] as the prediction. Both analyses assume that no locking attempt fails and comparisons are never stale. To avoid such issues, we execute the tests sequentially. Therefore, we parametrize the MultiQueue by the total number of PQs rather than of the queue factor in this section. Preliminary experiments with up to 128 threads showed that the effects of concurrency have a negligible impact on the quality.

Figure 7 shows the measured quality throughout the monotonic stress test with different numbers of candidate PQs d for the delete operation. When deleting from a randomly chosen PQ ($d = 1$), the rank error increases not only with the number of iterations but also with the number of elements in the PQ. In contrast, considering just two PQs ($d = 2$) stabilizes the rank error and makes it independent of the pre-fill. This result validates the fundamental design of the MultiQueue. As expected, the rank error further decreases for higher d . We set $d = 2$ for the rest of the experiments as it strikes a good balance between quality and cache efficiency.

Figure 8 shows the mean rank error for a varying number of PQs, and the distribution of rank errors for 256 PQs. The mean rank error is indeed linear in the number of PQs and the factor converges almost exactly to the predicted value of $\frac{5}{6}$, which is slightly worse than our estimation of $\frac{1}{2}$. The reason for the discrepancy for smaller numbers of PQs is likely due to an optimization in the MultiQueue that ensures that it picks two distinct PQs for deletion, whereas the prediction assumes that both PQs are chosen independently. For larger numbers of PQs, the optimization becomes less significant. The shape of the rank error distribution also closely matches the prediction, which is slightly worse than our estimation.

8.4 Parameter tuning

In this section, we explore the parameter space of the MultiQueue systematically and identify a set of good configurations. We generally optimize for good scalability and performance with many threads. We first tune the buffers and the k -ary heap, and then find combinations of stickiness period and queue factors that yield good throughput–quality trade-offs. We evaluate buffer sizes of $C \in \{0, 4, 16, 64, 256, 1024\}$ and heap arities of $k \in \{2, 4, 8, 16\}$. To reduce complexity, we only

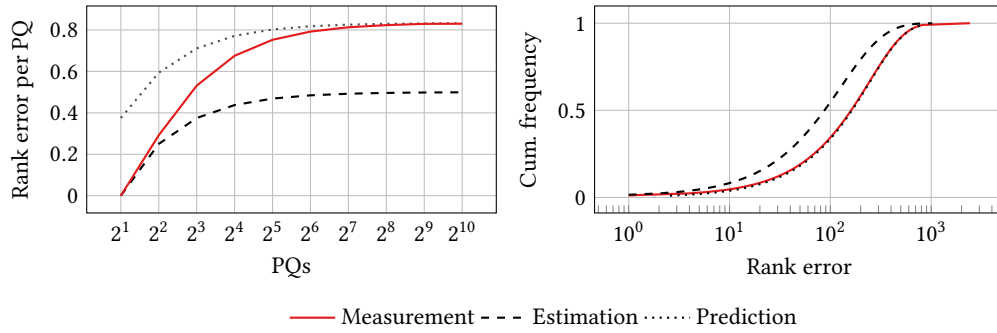


Fig. 8. Rank errors during the monotonic stress test with a pre-fill of $n = 2^{20}$. The left plot shows the mean rank error for different numbers of PQs. The right plot shows the cumulative frequency of rank errors less than or equal to the x-axis value with 256 PQs. For reference, our estimation from Section 5.2 and the results from Walzer and Williams [50] are also shown.

consider the same buffer size for both the insertion and deletion buffer. Figure 9 shows a clear correlation between buffer size, throughput, and cache misses. The correlation is in line with our expectations that the performance of the MultiQueue is limited by cache misses, and confirms that buffering is an effective mitigation measure. While larger buffers increase the overall throughput due to higher cache locality, they also increase the runtime of operations that need to flush or refill buffers. Higher variance in the runtime of operations leads to deteriorating quality, which can be observed with buffer sizes larger than $C = 16$. Buffer sizes larger than $C = 256$ yield only minor improvements in throughput while worsening the quality significantly. The heap arity only has a minor impact on the throughput and quality, but $k = 8$ seems to be the best choice overall. Our final configuration with $C = 16$ and $k = 8$ yields the highest throughput without sacrificing quality.

The queue factor c and the stickiness period s offer a trade-off between throughput and quality. Intuitively, the higher the number of PQs and the longer the stickiness period, the higher the throughput and the rank errors. Figure 10 confirms this intuition and shows that the quality deteriorates more quickly with increasing queue factors than with higher stickiness. Consequently, configurations with $c = 2$ dominate configurations with higher c and comparable throughput. The speedup from increasing the stickiness period eventually stagnates, presumably because the cache misses from switching PQs become insignificant. For higher throughput, the queue factor has to be increased at the cost of significantly worse quality. With higher queue factors, each PQ contains fewer elements, so the operations are faster and longer stickiness periods are possible before the throughput stagnates. Notably, for $c = 2$ and the simple stickiness variant, not only the throughput, but also the quality stagnates. A possible explanation is that switching PQs can trigger other threads to switch PQs as well, before their stickiness period is over. This cascading effect is less pronounced with higher queue factors, since it becomes more likely to switch to an unused PQ. The swap stickiness variant also mitigates the cascading effect and yields slightly higher throughput than the simple variant with comparable quality.

Table 1 shows configurations that yield interesting trade-offs between average rank error and throughput. For comparison with the competitors, we only consider configurations with $c = 2$, as other Pareto-optimal configurations with $c > 2$ exhibit significantly worse quality. The sharp increase in rank error for $c > 2$ is likely due to the longer time required for all PQs to be selected for deletion by some thread again, reminiscent of the coupon collector’s problem. Workloads in which not all threads frequently perform deletions (such as producer-consumer scenarios) may experience higher rank errors than in the monotonic stress test for the same configurations.

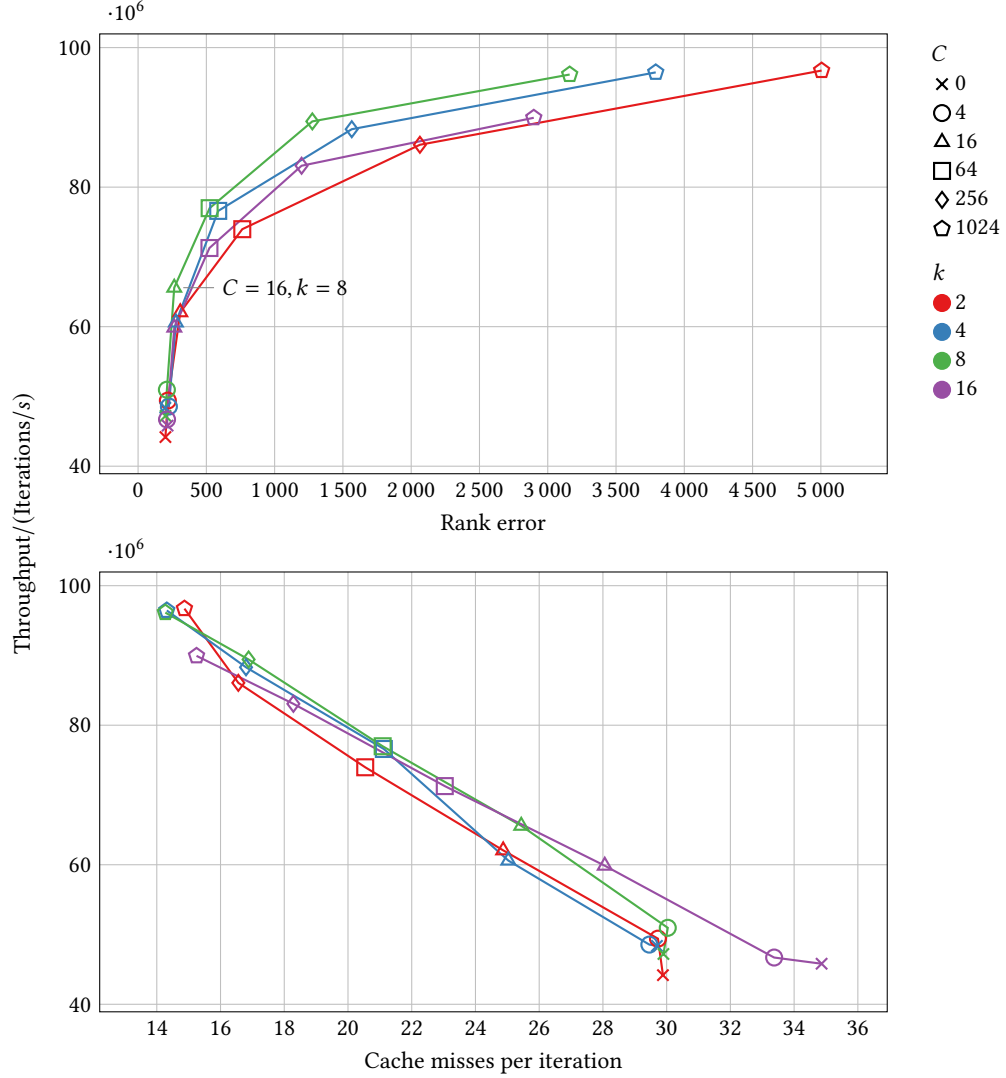


Fig. 9. Throughput versus rank error and L2 cache misses for varying buffer sizes C and heap arities k with 128 threads and $n = 2^{23}$ elements pre-fill. The buffer size is the same for the insertion buffer and the deletion buffer. The annotated data point corresponds to the final configuration with $C = 16$ and $k = 8$.

8.5 Comparison

We now compare the final configurations of the MultiQueue to state-of-the-art concurrent priority queues found in the literature. Besides relaxed PQs, we include linearizable PQs in order to assess the trade-off between quality and throughput in real-world applications. In particular, we include the linearizable PQ by Lindén and Jonsson [24], as it is a popular point of reference in the literature (e.g., [3, 8, 40, 41, 53]) While there is more recent work on linearizable PQs (e.g., [8, 40]) that outperforms the Linden PQ empirically by up to a factor of 5, we expect all linearizable PQs to have significantly worse scalability and throughput (with high thread counts) than relaxed PQs. As a baseline for the stress

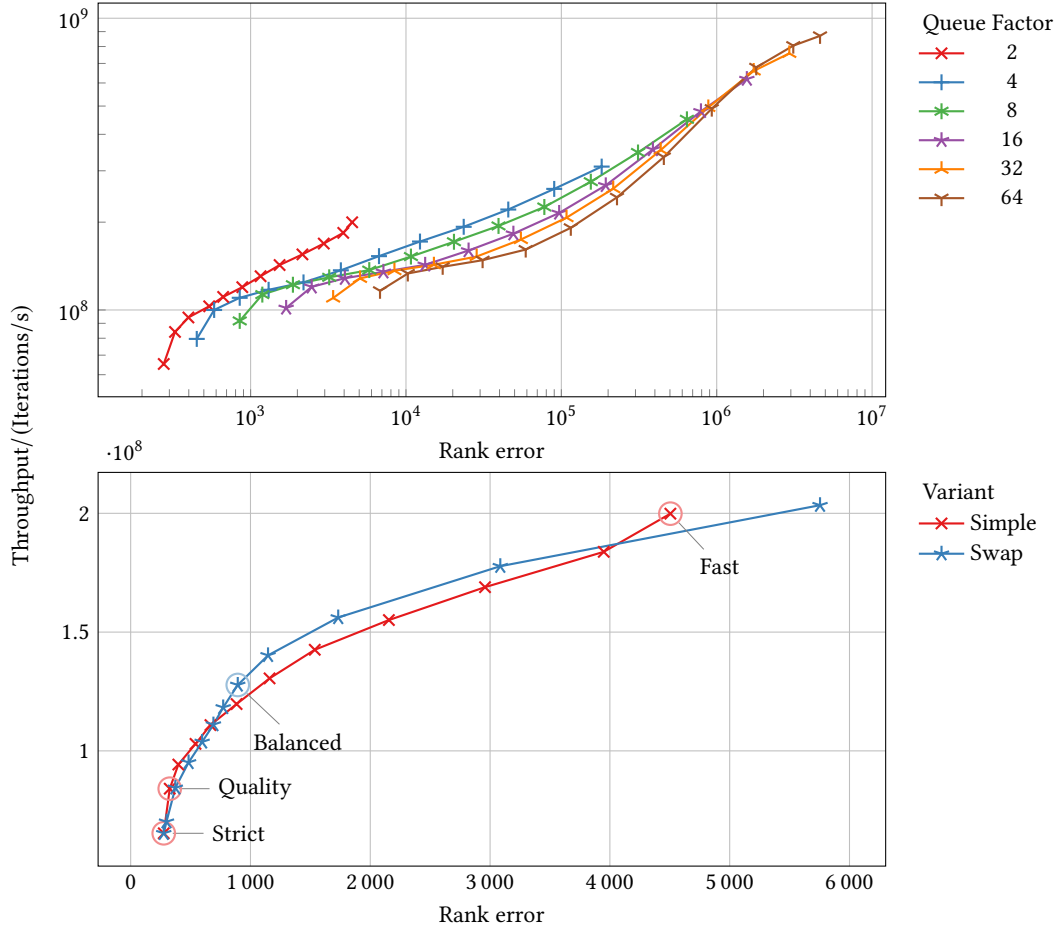


Fig. 10. The mean rank error and throughput with stickiness for different queue factors with 128 threads and a pre-fill of $n = 2^{23}$. The points are connected, starting with no stickiness, followed by increasing stickiness periods $2^2, 2^3, \dots, 2^{12}$, for better readability. The upper plot shows the impact of stickiness (simple variant) for different queue factors. The lower plot compares the simple and swap stickiness variants for $c = 2$.

Table 1. A selection of Pareto-optimal configurations.

Name	Stickiness	c	s	Throughput 10^6 Iterations/s	Rank error
Strict	No	2	–	65	276
Quality	Simple	2	4	84	326
Balanced	Swap	2	256	128	894
Fast	Simple	2	4096	200	4505
	Simple	4	4096	310	181 811
	Simple	16	4096	619	1 563 132
	Simple	64	4096	870	4 623 838

tests, we also include the buffered k -ary heap used by the MultiQueue, guarding each access with a mutual exclusion lock. The following list gives an overview of the competitors. Detailed descriptions of most competitors can be found in Section 3.

MQ strict, quality, balanced, fast MultiQueue configurations from Table 1.

k -LSM 4, 256, 1024, 4096 The k -LSM by Wimmer et al. [53] with $k = 4$, $k = 256$, $k = 1024$ and $k = 4096$, respectively.

CA-PQ The contention-avoiding priority queue by Sagonas and Winblad [41].

Spraylist The Spraylist by Alistarh et al. [3] in the default configuration.

SMQ Standalone adaption of the stealing MultiQueue implementation¹² by Postnikova et al. [34]. We use a stealing probability of $\beta = \frac{1}{8}$ and a batch size of $b = 64$ as suggested by the authors.

Linden Linearizable PQ by Lindén and Jonsson [24].

TBB PQ Linearizable Concurrent priority queue from Intel® oneAPI Threading Building Blocks.¹³

Locked Heap Buffered k -ary heap used by the MultiQueue with a mutual exclusion lock for each operation. This competitor is included to provide a linearizable baseline for the stress tests.

The implementations of the CA-PQ, the Spraylist and the Linden PQ are taken from the k -LSM repository¹⁴. Unfortunately, the implementations of the k -LSM, the CA-PQ, the Spraylist and the Linden PQ are not compatible with the ARM AAarch64 ISA; they either fail to compile or behave incorrectly due to the more relaxed memory ordering compared to x64-64.

8.5.1 Stress tests. Figure 11 plots the quality against the throughput for all competitors on the monotonic stress test with two different pre-fills on machine AMD. The full results with a larger spectrum of thread counts and pre-fills are given in Appendix A. The MultiQueue configurations dominate the Pareto-front except for very small rank errors with all tested pre-fills. When comparing with competitors that achieve similar rank errors on 128 threads, the advantages range from $4.5\times$ to $25\times$ in throughput. With large pre-fills, the MultiQueue achieves significantly higher throughput than all other competitors. The k -LSM 1024 and the k -LSM 4096 achieve similar throughput as the fast MultiQueue with 128 threads for the smallest pre-fill but exhibit an order of magnitude higher rank errors. On large pre-fills, the k -LSM fails to scale and is slower than the MultiQueue by almost two orders of magnitude with 128 threads. The CA-PQ achieves competitive throughput with the MultiQueue but fails to scale beyond 32 threads on small pre-fills. On large pre-fills, the CA-PQ exhibits similar quality to the balanced MultiQueue but is 5 times slower. The SMQ offers competitive throughput to the MultiQueue on small pre-fills, but the quality is at least an order of magnitude worse in most cases. Moreover, the quality of the SMQ deteriorates with the pre-fill, and it exhibits the highest rank errors of all competitors on large pre-fills. The Spraylist is robust to the pre-fill and exhibits rank errors very similar to the balanced MultiQueue but is an order of magnitude slower. Unsurprisingly, the linearizable PQs and the k -LSM 4 yield very low rank errors but do not offer competitive throughput, especially with higher thread counts.

Figure 12 shows the corresponding throughput scaling behaviour of all competitors on all three machines using the same pre-fills as in Figure 11. For better readability, only the balanced MultiQueue configuration and the best (highest maximum throughput) k -LSM variant are shown. The full results are given in Appendix B. The MultiQueue variants generally scale well, even when using hardware multithreading. The performance dips slightly when going upwards in

¹²<https://github.com/npostnikova/mq-based-schedulers>

¹³The concurrent priority queue from Intel® oneAPI Threading Building Blocks.

¹⁴<https://github.com/klsmq/klsm>

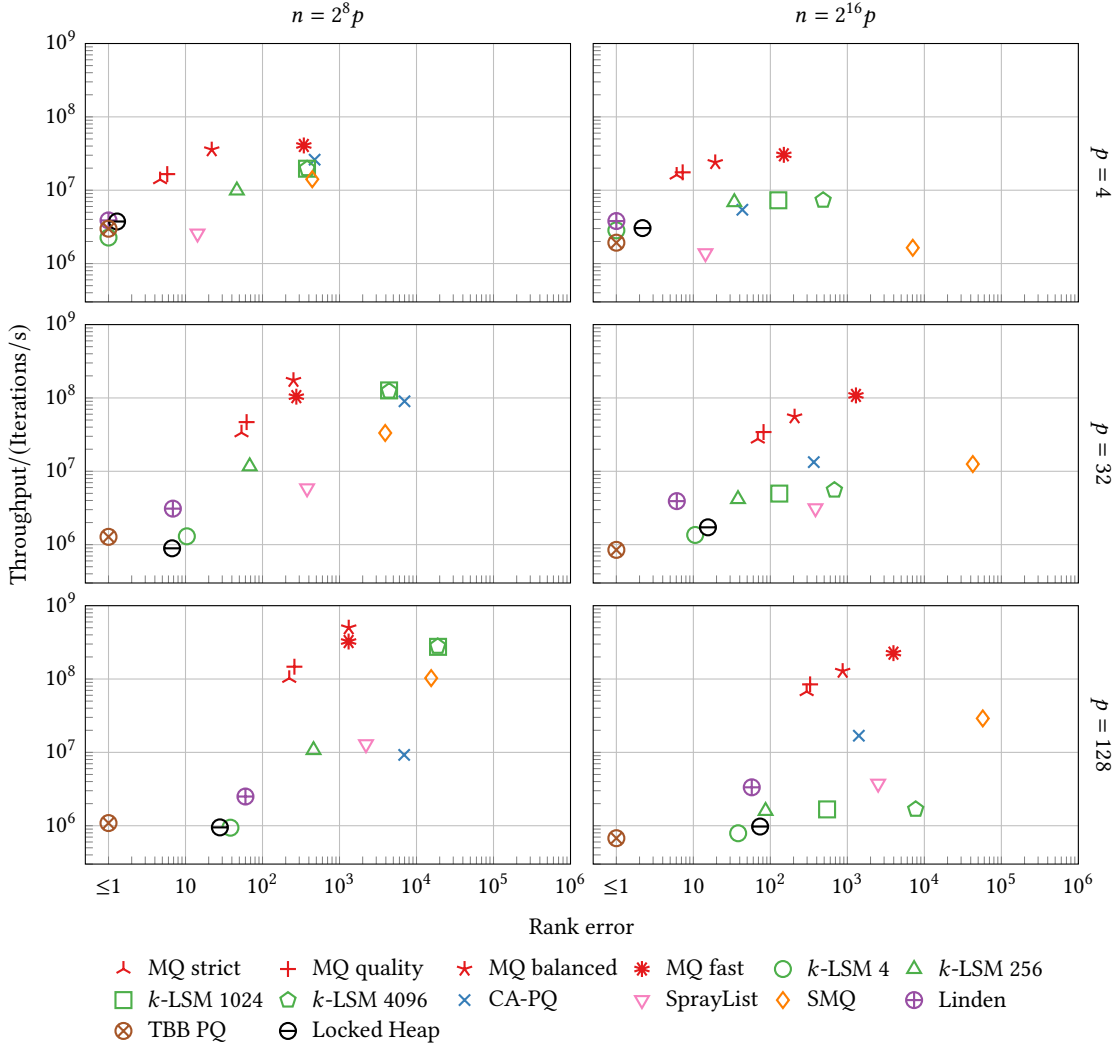


Fig. 11. Mean rank error versus throughput for the monotonic stress test with different pre-fills n on machine AMD. Each row of plots corresponds to a different thread count (p) and each column to a different pre-fill (n). The label “ ≤ 1 ” indicates that the rank error is less than or equal to 1.

the architectural hierarchy. This effect is most pronounced on the Intel-NUMA machine when going from two to four sockets. However, this is a price one is often willing to pay in order to be able to globally coordinate multiple threads by priority driven scheduling of work—the overall application may still profit from being able to use more threads.

As a close relative to the MultiQueue, the SMQ has similar scaling behavior, albeit with lower absolute performance. The CA-PQ scales up to a moderate number of threads but then experiences sharp drops in performance, likely when it has to fallback to the global skip list. We view it as likely that adapting some tuning parameters might be able to remedy this problem, albeit at the price of even higher rank errors. The more relaxed variants of the k -LSM scale well

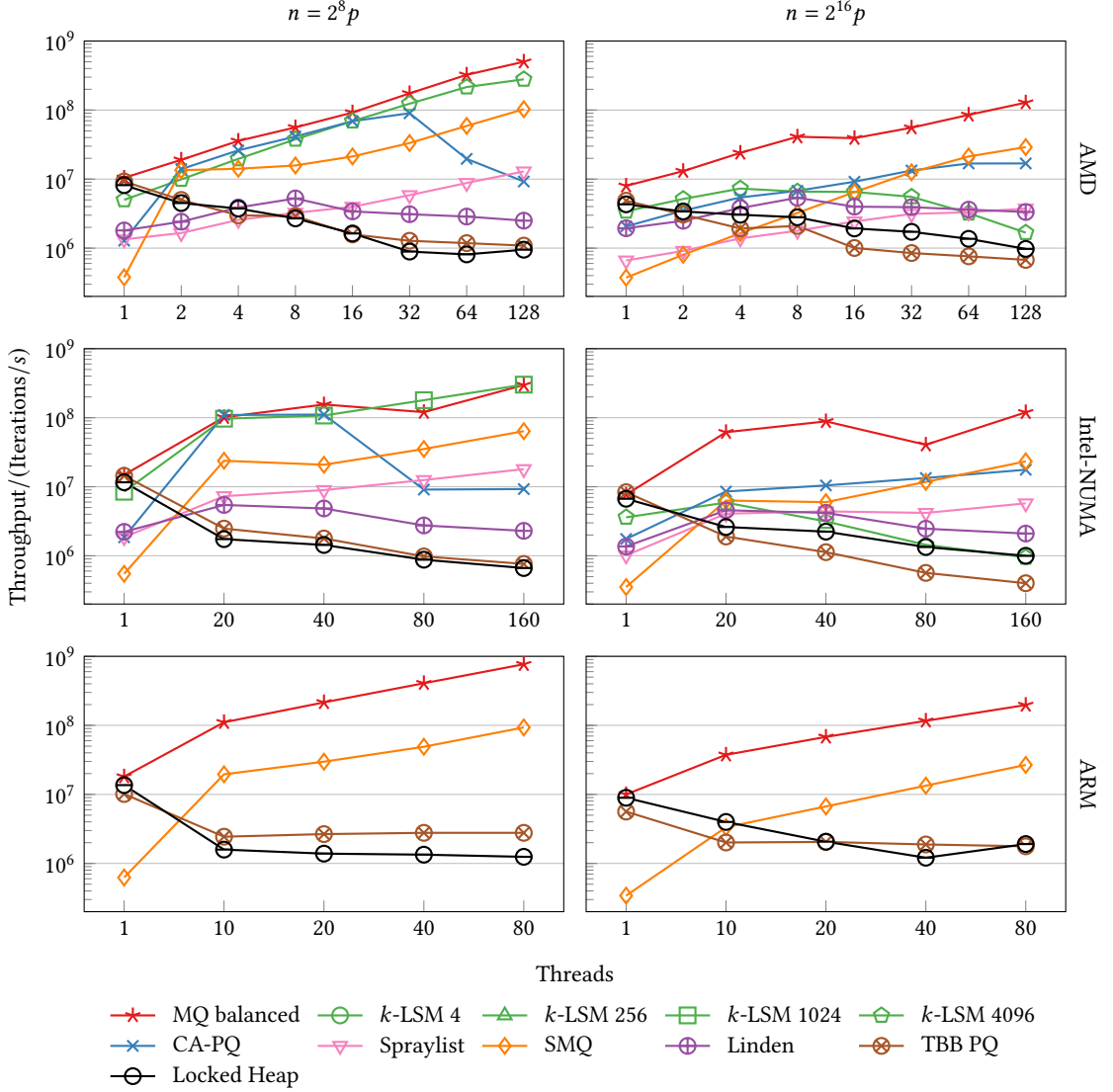


Fig. 12. Weak scaling experiment showing the throughput for the monotonic stress test versus the number of threads on different machines. Each row corresponds to a different machine and each column to a different pre-fill. Only the k -LSM variant with the highest throughput is shown for better readability. Missing plots on the ARM machine are due to incompatibility with the AArch64 ISA.

for small queues, but the scalability is poor in all other configurations. We attribute the particularly bad performance for large queues to the bottleneck introduced by the expensive merging within LSM trees. As expected, the linearizable queues do not scale at all. Throughput even *decreases* by an order of magnitude when going to the maximum number of threads, and can be up to 300 times slower than the fastest relaxed queues on the AMD machine. This underlines the utility of relaxed queues.

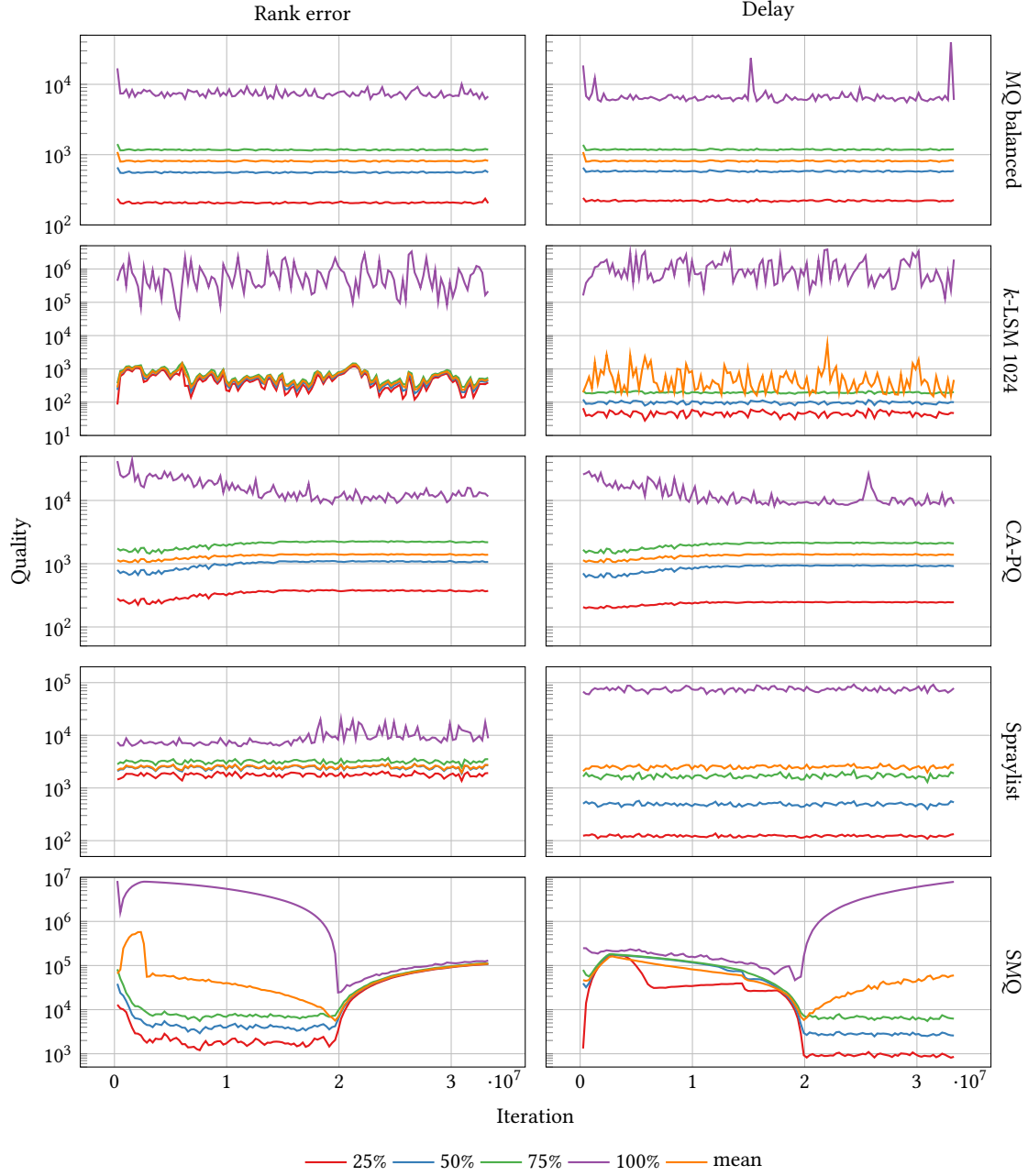


Fig. 13. Development of the rank errors and delays of the monotonic stress test with a pre-fill of $n = 2^{16}p$ and 128 threads on machine AMD. The iterations are grouped into bins of size 2^{18} , and for each bin the 25%, 50%, 75% and 100% quantiles as well as the mean are plotted. Each row corresponds to one competitor and the two columns show the rank errors and delays, respectively.

Examining the rank errors and delays in more detail, Figure 13 shows their development over time for a selection of the relaxed PQs. The MultiQueue exhibits very stable rank errors and delays over time, and both metrics behave very similarly. The 25% and 75% quantiles of the rank errors and delays are close together, indicating that most rank errors and delays are within a small range. The largest measured rank errors and delays are roughly one order of magnitude larger than the mean. The rank errors and delays fluctuate much more for the k -LSM than for the other queues. These fluctuations reach an order of magnitude for the maxima and also exceed the claimed rank error bounds by an order of magnitude. This is likely due to occasional expensive sequential merge operations of large parts of the log-structured merge tree during which no elements can be deleted from the merged lists. While the rank error quantiles are very close together, the delay quantiles are more spread out. The CA-PQ behaves similarly to the MultiQueue but exhibits a slight increase in mean rank errors and delays over time, while the maximum rank errors and delays decrease. The Spraylist exhibits a very small range of rank errors but delays are widely spread. The SMQ shows very inconsistent behavior over time, which is unexpected for such a homogeneous benchmark, and given the consistent behavior of the other competitors. Due to these inconsistencies, the overall mean rank errors and delays vary significantly with the number of iterations for the SMQ.

Figure 14 shows the throughput as the average time per operation for the insert-delete stress test on all machines. Again, we only show the balanced MultiQueue configuration and the best k -LSM variant for better readability. The full results are given in Appendix C. Most competitors, with the exception of the CA-PQ, finish the insertions faster than the deletions. Only the SMQ can compete with the insertion times of the MultiQueue; however, it scales worse with the number of threads. All other competitors are at least an order of magnitude slower than the MultiQueue and SMQ for high thread counts. The CA-PQ generally has the fastest deletions for low thread counts but is slower than the MultiQueue for high thread counts.

8.5.2 Single-Source Shortest-Path. The single-source shortest-path (SSSP) problem is a fundamental and widely known graph problems. Given a weighted graph G and a source node s in G , the goal is to find the shortest path from s to all other nodes in G . The most famous algorithm for the SSSP problem with non-negative edge weights is probably Dijkstra’s algorithm [13]. Hence, a natural parallel algorithm is a modified version of Dijkstra’s algorithm using a concurrent PQ.¹⁵ The algorithm maintains an array of tentative distances to each node and a PQ that stores nodes ordered by their tentative distances. Initially, the tentative distance is 0 for s and ∞ for all other nodes, with s being the only node in the PQ. Then, all threads repeatedly process nodes from the PQ. If a node’s tentative distance remains unchanged since its insertion, the node is *scanned*; otherwise, it is discarded. Scanning involves checking each outgoing edge to determine whether it yields a shorter tentative distance to a neighboring node. If so, the neighbor’s tentative distance is updated atomically in the array, and the neighbor is (re)inserted into the PQ. Note that nodes are reinserted into the PQ with improved tentative distances instead of being updated in-place since most implementations of relaxed PQs do not support these updates. The algorithm terminates when the PQ is empty and no thread is currently scanning a node. Since the total number of operations is unknown in advance, we employ the termination detection algorithm shown in Figure 6. The delete operations of the k -LSM and Spraylist implementations sometimes fail unexpectedly, violating the assumptions of the termination detection algorithm that the PQ is guaranteed to be empty if the last operation of every thread was a failed deletion and currently no insertions are in progress.¹⁶ We use the following approach to adapt the termination detection algorithm for these competitors. Each thread additionally counts its

¹⁵Since we use the algorithm as a benchmark for PQs, we do not prioritize the competitiveness of the algorithm itself. Several specialized parallel algorithms for the SSSP exist (e.g., Δ -stepping [30]).

¹⁶We do not see fundamental reasons why these PQs could not be implemented in a way that satisfies this assumption.

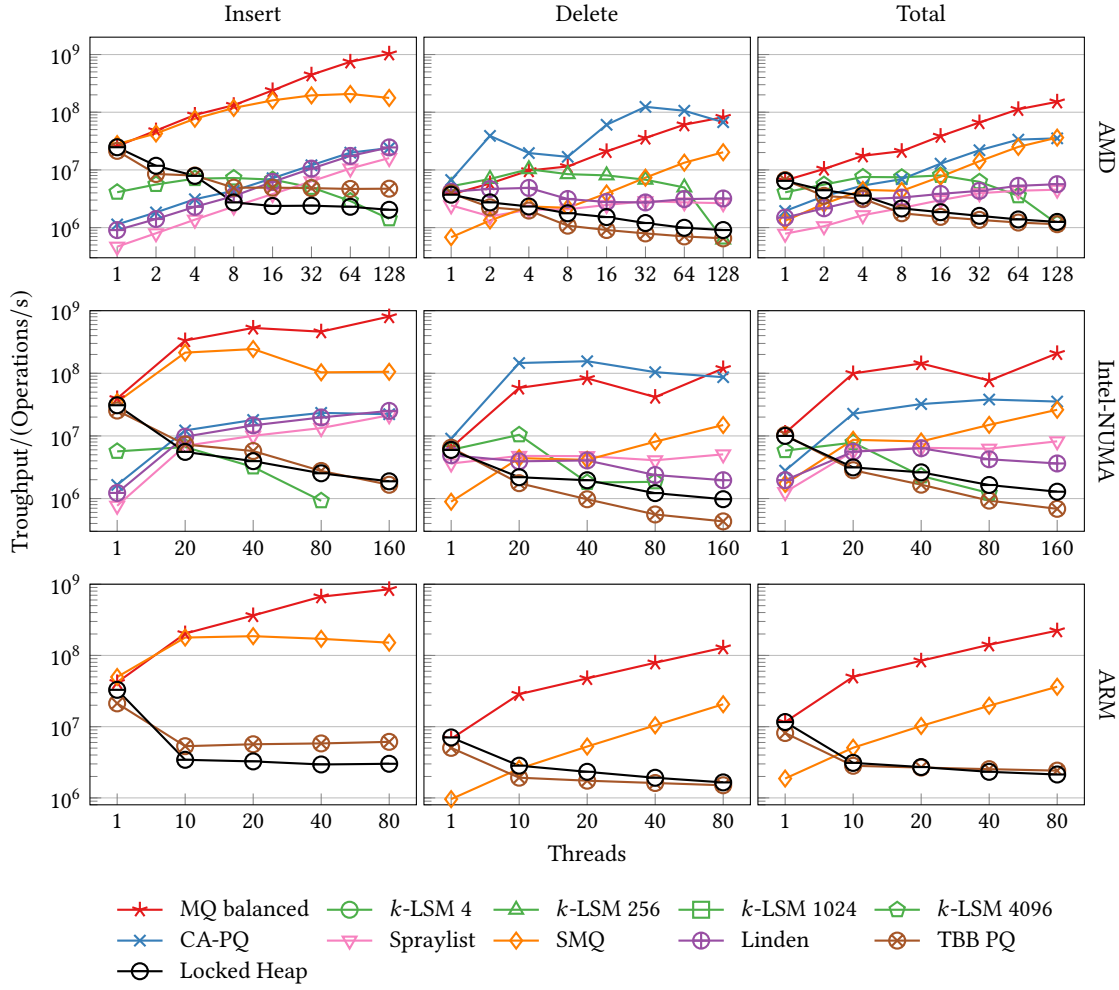


Fig. 14. Throughput comparison for the insert-delete stress test with 2^{20} elements per thread on different machines. Only the k -LSM variant with the highest throughput is shown for better readability. Data points for the k -LSM variants on machine Intel-NUMA for 160 threads are missing due to crashes. Missing lines in the ARM plot are due to incompatibility with the ARM AArch64 ISA.

insertions and deletions locally. Instead of terminating when all threads are idle, they synchronize to sum up all counters. If the total number of insertions is equal to the total number of deletions, all threads can terminate. Otherwise, the threads are reset to the working state and continue processing elements.

Nodes can be deleted and scanned before all nodes with shorter tentative distances are fully scanned, meaning shorter paths to a node may still be found even after it has been scanned. This results in redundant scans, making the algorithm *label-correcting*, even with linearizable PQs. In contrast, classical Dijkstra’s algorithm is *label-setting*—each node is only scanned once and its distance does not change afterward. The number of redundant scans depends on the quality of the PQ, but in the worst case, a node might be scanned an exponential number of times relative to the number of nodes in the graph [48].

Table 2. Graphs used for the SSSP problem.

Name	Nodes/ 10^6	Edges/ 10^6	Max. degree
NY	0.3	0.7	8
CAL	1.9	4.7	8
CTR	14.1	34.3	9
GER	20.7	41.8	9
USA	23.9	58.3	9
RHG ₂₀	1	10.3	90 391
RHG ₂₂	4.2	41.3	940 912
RHG ₂₄	16.8	159.7	594 914

We evaluate the SSSP problem on road networks and random hyperbolic graphs (RHGs), and report the solving time for different thread counts as well as the number of scanned nodes. The edge weights in the road networks correspond to an estimated travel time, and the weights in the hyperbolic graphs correspond to the integer approximation of the hyperbolic distance. All road networks were obtained from the 9th DIMACS implementation challenge¹⁷, except for the GER graph¹⁸. We used KaGen [16] to generate hyperbolic graphs with 2^{20} , 2^{22} and 2^{24} nodes, a configured average degree of 16 and a power law exponent of $\gamma = 2.3$. An overview of the graphs is given in Table 2.

Figure 15 shows the time to solve the SSSP problem on these graphs with varying numbers of threads. Again, we only show the balanced MultiQueue configuration and the fastest k -LSM variant for better readability. The full results are given in Appendix D. The MultiQueue and the SMQ generally scale well with the number of threads. The MultiQueue yields faster solving times than all other competitors on all graphs, achieving speedups of 50 with 128 threads over the sequential algorithm. These results indicate that it might be interesting to investigate a relaxed version of Dijkstra’s algorithm as a competitive parallel SSSP solver. The SMQ offers similar performance to the MultiQueue on the hyperbolic graphs but is slightly slower on the road networks. While the k -LSM has good speedups on the larger road networks, it does not scale at all on the random hyperbolic graphs. This is consistent with the observations from our stress tests, where we observed that the k -LSM performs best when the queues are small—road networks have high diameter and thus small average search frontiers while the search frontier in RHGs quickly expands to encompass a large part of the graph. Notably, the k -LSM 1024 performs better than the more relaxed variants on the road networks. On the RHG₂₀ graph, the SMQ is faster than the MultiQueue for 128 threads, since the MultiQueue fails to scale beyond 32 threads. However, the strict and quality variants of the MultiQueue still outperform the SMQ (refer to Appendix D). On the RHG graphs, the fast MultiQueue does not scale beyond 16 threads and becomes slower than the other MultiQueue configurations.

To demonstrate the effects of relaxation on performance, Figure 16 shows the excess work done by the best performing PQs. Besides the much slower k -LSM 256, the MultiQueue variants process the fewest nodes on the road networks, which is consistent with the solving times. The k -LSM 4096 processes the most nodes on road networks (except for the NY graph), up to $18\times$ more than the sequential algorithm and three times as many as the k -LSM 1024. This explains the better performance of the k -LSM variants with smaller k . The amount of excess work done on the hyperbolic graphs is generally much lower than on the road networks. A possible reason is that the much lower diameter of hyperbolic graphs leads to much larger average queue size. In comparison, the rank errors are very small. In other words, the shortest path algorithm has plenty of nodes in the queue to choose from, and the nearest of those nodes

¹⁷<http://www.diag.uniroma1.it/challenge9/download.shtml>

¹⁸<https://i11www.iti.kit.edu/resources/roadgraphs.php>

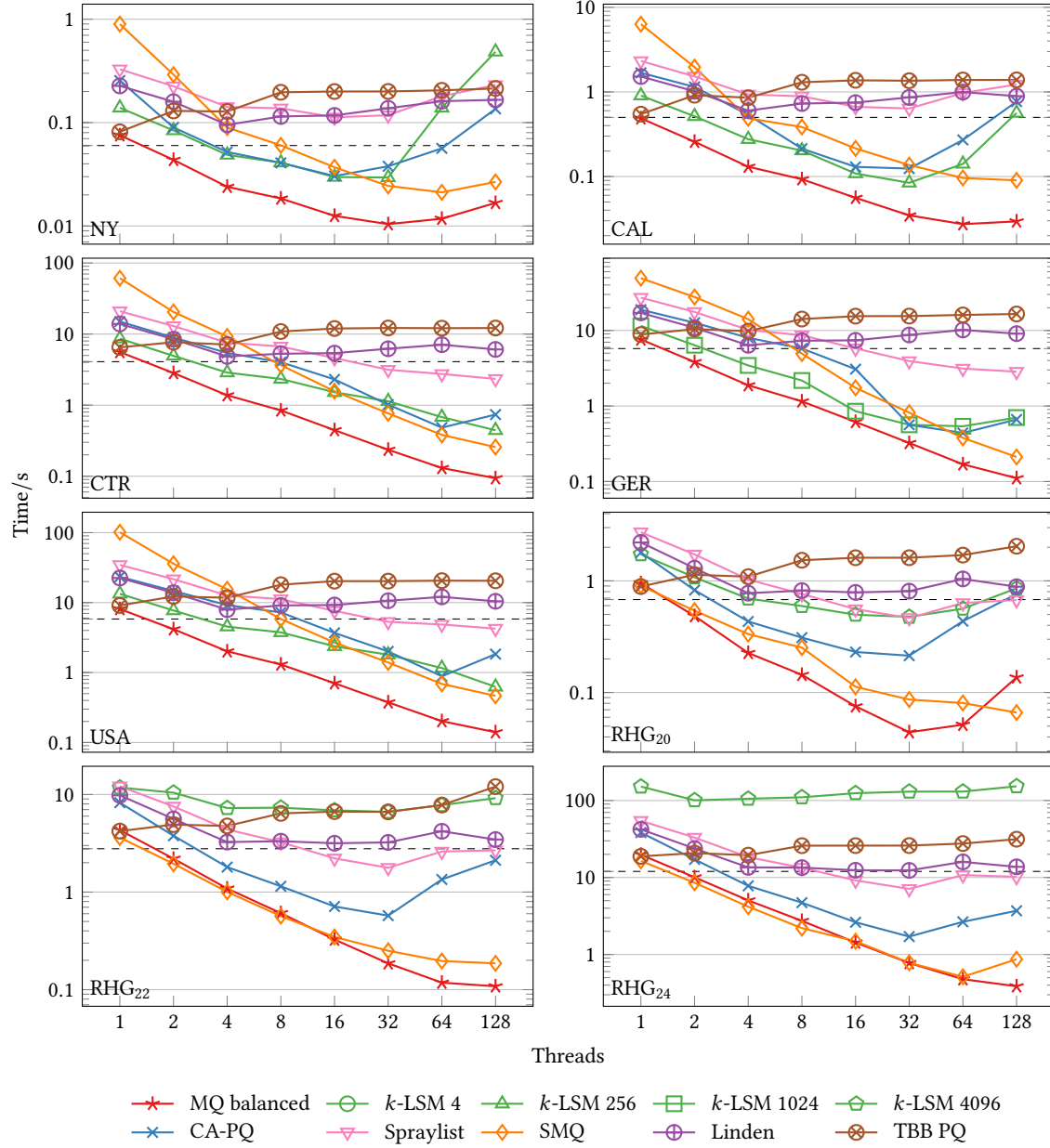


Fig. 15. Time to solve the SSSP problem with different numbers of threads on machine AMD. The black dashed line is the solving time of the sequential algorithm. Only the k -LSM variant with the fastest solving time is shown for better readability.

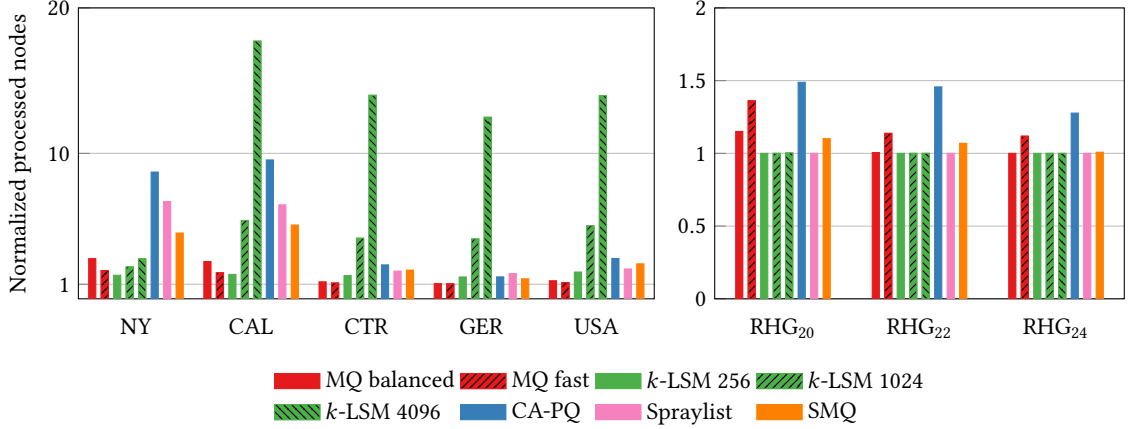


Fig. 16. Normalized processed nodes with 128 threads with respect to the sequential SSSP algorithm. The normalization is with respect to the number of processed nodes by the sequential algorithm. For clarity, only the best performing priority queues are shown. The k -LSM 256 and 1024 do not finish within the time limit of 300 s on graph RHG₂₄.

are likely to have already reached their final shortest path distance. While the CA-PQ exhibits the most overhead, the increased overhead of the fast MultiQueue compared to the balanced MultiQueue seems to be responsible for the worse performance on the RHG graphs.

In summary, relaxation is beneficial for the concurrent version of Dijkstra’s algorithm, but the amount of relaxation has to be chosen carefully to avoid excessive node scans.

8.5.3 Knapsack. Lastly, we evaluate the performance of the competitors on a branch-and-bound algorithm for the knapsack problem. The PQ stores partial solutions, each containing the current value, weight, and index up to which items have been considered. The lower bound of a partial solution is given by greedily adding items to the knapsack in order of value density until the next item no longer fits or no items remain. To speed up this step, we use a *finger search*, starting the search for the next item at the index of the last added item. The upper bound is computed by adding the fractional portion of the next item to completely fill the knapsack. These bounds are used to prune solutions that cannot be optimal. The partial solutions in the PQ are ordered in decreasing order of the upper bound. Similar to the SSSP problem, the algorithm terminates when the PQ is empty, and the number of nodes to process is unknown in advance. Hence, we use the same termination detection algorithm as for the SSSP problem.

We generate instances following a standard approach [26, Section 2.10], adapted to integer weights and values; see also [43]. For each item, the weight w is drawn uniformly at random from $[1, W]$, and the value v is given by $v = w + k$, where k is drawn uniformly at random from $[V, fV]$ for $W \in \{1000, 10000\}$, $V = \frac{W}{10}$, $f \in \{1.1, 1.15, 1.2\}$. The target capacity is 50% or 90% of the total weight of all items. These parameters are chosen to generate “interesting” instances that are nontrivial to solve yet allow for significant pruning opportunities. For each parameter combination, we generate 100 instances with 1000 items. From these, we select instances that the sequential algorithm solves within 1–15 s, resulting in a final set of 147 instances. Each competitor is given a time limit of 60 s, leading to a total thread-time with 128 threads that is 512× that of the sequential algorithm.

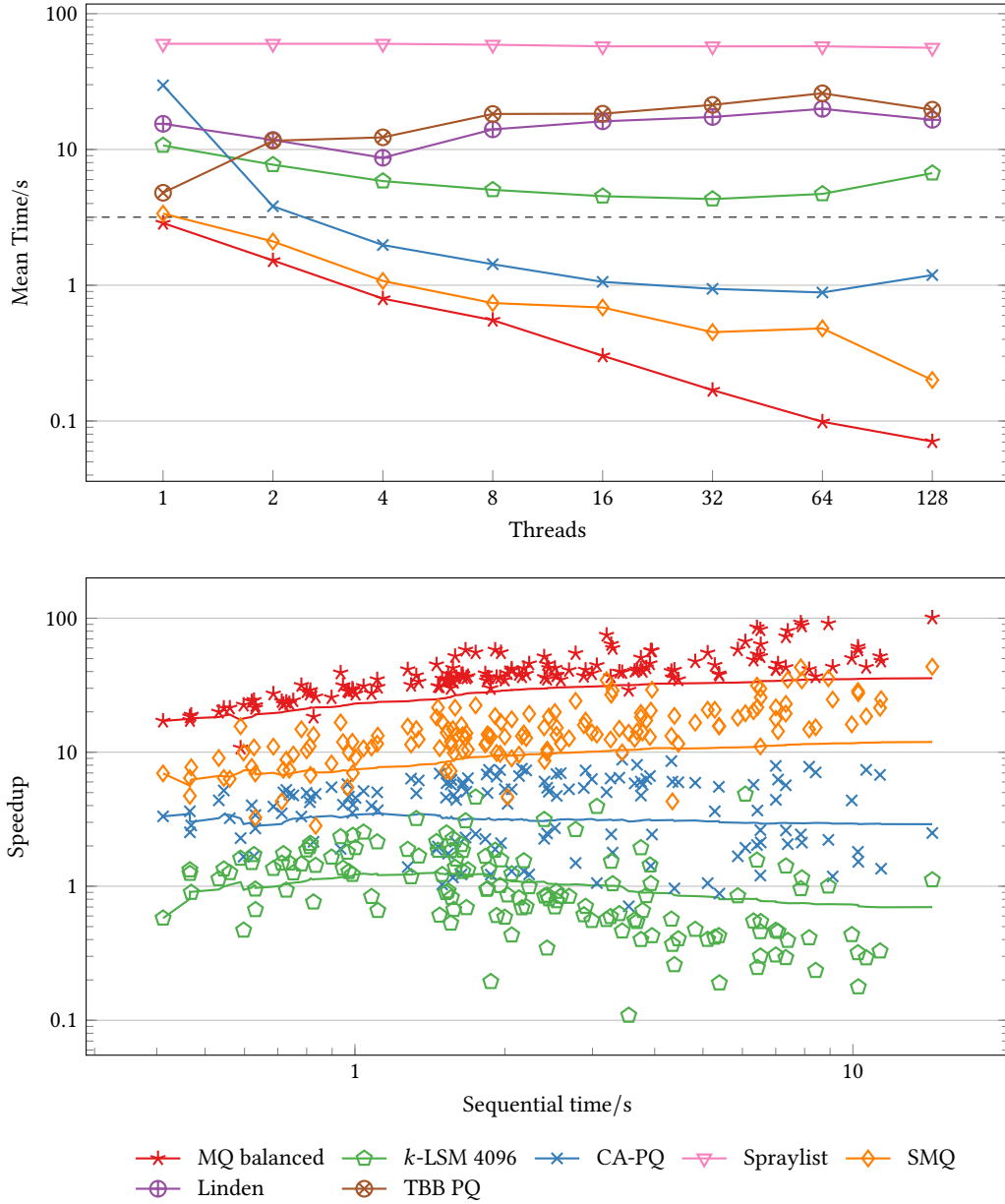


Fig. 17. Top: Arithmetic mean of times to solve each instance of the knapsack problem with different thread counts on machine AMD. If an instance is not solved within the time limit of 60 s, the solving time is set to the time limit. The dashed black line indicates the sequential solving time. Bottom: Speedup with 128 threads for each instance, ordered by the sequential solving time. The lines are the cumulative harmonic mean (from left to right) of the speedup. Only the best performing relaxed priority queues are shown for clarity.

Table 3. Processed nodes and speedup for the knapsack problem with 128 threads. r is the number of processed nodes, normalized to the number of processed nodes by the sequential algorithm. The speedup is computed relative to the sequential algorithm.

PQ	Geom. mean r	Max r	Harm. mean speedup	Max. speedup
MQ balanced	1.08	2.5	35.65	100.84
k -LSM 4096	1.12	2.47	0.7	4.85
CA-PQ	1.27	4.59	2.9	8.8
SMQ	1.11	4.71	11.92	43.45

Figure 17 shows the average solving times and speedups per instance for the competitors. For better readability, we only show the balanced MultiQueue configuration and the fastest k -LSM variant. The full results are given in Appendix E. Table 3 shows the aggregated number of processed nodes and speedups for the best performing PQs.

The MultiQueue performs best on average for all thread counts, followed by the SMQ. This is reflected in the individual speedups, as the MultiQueue achieves the highest speedups with 128 threads across almost all instances, averaging 35 and reaching up to 100 on some instances. The CA-PQ scales well up to 16 threads, after which its performance plateaus. Even the most relaxed k -LSM variant barely scales—presumably due to large PQ sizes—and remains slower than the sequential algorithm for all thread counts. Similarly, the linearizable PQs do not scale and are slower than the sequential algorithm. Unfortunately, the Spraylist fails to solve most instances within this time limit of 300 s. The best-performing PQs exhibit an overhead of about 10% in the number of processed nodes, demonstrating that relaxation is an effective way to speed up branch-and-bound algorithms. They are multiple orders of magnitude faster than the linearizable competitors and scale well.

9 Conclusion and Future Work

The MultiQueue demonstrates that the *two-choices paradigm* is a powerful concept for designing relaxed concurrent priority queues with good scalability and accuracy. Buffering and, more importantly, stickiness help to mitigate poor cache locality and reduce cache-coherence traffic.

Compared to linearizable queues, we observed several orders of magnitude better throughput. Thus, applications that profit from fine-grained priority-based scheduling of tasks can be effectively parallelized using the MultiQueue. Still, benchmarking additional workloads, such as producer-consumer scenarios, and more complex applications might be interesting for better understanding the capabilities of the MultiQueue.

Although we have focused on locked k -ary heaps, the MultiQueue can be combined with a wide range of internal queues for various effects. Exploiting integer keys or relaxing the internal queues to simple bucket queues can further accelerate operation, e.g., see [55]. Using lock-free internal queues (and buffers) renders the MultiQueue deterministically lock-free in addition to being probabilistically wait-free.

Our (over)simplified analysis goes some way to explaining the good performance and quality of the MultiQueue. Although a tighter analysis of the basic (non-sticky) MultiQueue is now available [50], several questions remain open: What quality bounds can be proven when lifting all simplifying assumptions (hopefully, the result from [50] still applies), and what is the effect of the stickiness parameter s (one would expect a linear dependence of the rank error).

Our analysis of branch-and-bound algorithms exemplifies that, orthogonal to the inner workings of the queues, the concept of rank error and delay helps us to analyze specific applications. It seems interesting to extend this to further applications, e.g., shortest path computations, discrete event simulations, or computing Huffman codes.

The MultiQueue design we present in this work has some limitations that could be addressed in future work. The quality of the MultiQueue is not adaptive to the workload and its performance degrades with very few elements in the PQs. Further, the MultiQueue is limited to a fixed number of internal PQs, thus limiting the maximum number of threads that can concurrently access the PQ. A promising first step to address these limitations would be to adapt the stickiness period or the number of deletion candidates dynamically depending on the workload and contention level, similar to the CA-PQ. Dynamically adjusting the number of PQs is more challenging but still an interesting direction to explore.

The conceptual simplicity and good scalability of the MultiQueue also make it an attractive candidate for a GPU implementation. Despite current GPUs mostly use batch parallelism, it would be interesting to see whether a design based on the MultiQueue might be useful for asynchronous, massively parallel, priority-driven computing on GPUs.

References

- [1] Dan Alistarh, Trevor Brown, Justin Kopinsky, Jerry Z. Li, and Giorgi Nadiradze. 2018. Distributionally Linearizable Data Structures. In *30th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*. <https://doi.org/10.1145/3210377.3210411>
- [2] Dan Alistarh, Justin Kopinsky, Jerry Li, and Giorgi Nadiradze. 2017. The Power of Choice in Priority Scheduling. In *ACM Symposium on Principles of Distributed Computing (PODC)*. <https://doi.org/10.1145/3087801.3087810>
- [3] Dan Alistarh, Justin Kopinsky, Jerry Li, and Nir Shavit. 2015. The SprayList: A Scalable Relaxed Priority Queue. In *20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. <https://doi.org/10.1145/2688500.2688523>
- [4] Hagit Attiya, Rachid Guerraoui, Danny Hendler, Petr Kuznetsov, Maged M. Michael, and Martin Vechev. 2011. Laws of Order: Expensive Synchronization in Concurrent Algorithms Cannot Be Eliminated. *ACM SIGPLAN Notices* 46, 1 (Jan. 2011). <https://doi.org/10.1145/1925844.1926442>
- [5] Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. 1999. Balanced Allocations. *SIAM J. Comput.* 29, 1 (Jan. 1999). <https://doi.org/10.1137/S0097539795288490>
- [6] Petra Berenbrink, Artur Czumaj, Angelika Steger, and Berthold Vöcking. 2006. Balanced Allocations: The Heavily Loaded Case. *SIAM J. Comput.* 35, 6 (Jan. 2006). <https://doi.org/10.1137/S009753970444435X>
- [7] Timo Bingmann. 2018. TLX: Collection of Sophisticated C++ Data Structures, Algorithms, and Miscellaneous Helpers. <https://panthema.net/tlx>, retrieved Oct. 7, 2020.
- [8] Anastasia Braginsky, Nachshon Cohen, and Erez Petrank. 2016. CBPQ: High Performance Lock-Free Priority Queue. In *Euro-Par 2016: Parallel Processing*. 460–474. https://doi.org/10.1007/978-3-319-43659-3_34
- [9] Irina Calciu, Hammurabi Mendes, and Maurice Herlihy. 2014. The Adaptive Priority Queue with Elimination and Combining. In *28th Symposium on Distributed Computing (DC)*. https://doi.org/10.1007/978-3-662-45174-8_28
- [10] Narsingh Deo and Sushil Prasad. 1992. Parallel Heap: An Optimal Parallel Priority Queue. *The Journal of Supercomputing* 6, 1 (March 1992). <https://doi.org/10.1007/BF00128644>
- [11] Laxman Dhulipala, Guy Blelloch, and Julian Shun. 2017. Julianne: A Framework for Parallel Graph Algorithms Using Work-efficient Bucketing. In *29th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*. <https://doi.org/10.1145/3087556.3087580>
- [12] Martin Dietzfelbinger and Friedhelm Meyer auf der Heide. 1993. Simple, Efficient Shared Memory Simulations. In *5th ACM Symposium on Parallel Algorithms and Architectures (SPAA)*. <https://doi.org/10.1145/165231.165246>
- [13] E. W. Dijkstra. 2022. A Note on Two Problems in Connexion with Graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy* (1 ed.). Vol. 45.
- [14] Faith Ellen, Danny Hendler, and Nir Shavit. 2012. On the Inherent Sequentiality of Concurrent Objects. *SIAM J. Comput.* 41, 3 (2012). <https://doi.org/10.1137/08072646X> arXiv:<https://doi.org/10.1137/08072646X>
- [15] Faith Ellen, Yossi Lev, Victor Luchangco, and Mark Moir. 2007. SNZI: Scalable NonZero Indicators. In *ACM Symposium on Principles of Distributed Computing*. 13–22. <https://doi.org/10.1145/1281100.1281106>
- [16] Daniel Funke, Sebastian Lamm, Ulrich Meyer, Manuel Penschuck, Peter Sanders, Christian Schulz, Darren Strash, and Moritz von Looz. 2019. Communication-Free Massively Distributed Graph Generation. *J. Parallel and Distrib. Comput.* 131 (Sept. 2019). <https://doi.org/10.1016/j.jpdc.2019.03.011>
- [17] Andreas Haas, Michael Lippautz, Thomas A. Henzinger, Hannes Payer, Ana Sokolova, Christoph M. Kirsch, and Ali Sezgin. 2013. Distributed Queues in Shared Memory: Multicore Performance and Scalability through Quantitative Relaxation. In *ACM International Conference on Computing Frontiers (CF)*. <https://doi.org/10.1145/2482767.2482789>
- [18] Thomas A. Henzinger, Christoph M. Kirsch, Hannes Payer, Ali Sezgin, and Ana Sokolova. 2013. Quantitative Relaxation of Concurrent Data Structures. *ACM SIGPLAN Notices* 48, 1 (Jan. 2013). <https://doi.org/10.1145/2480359.2429109>
- [19] Maurice Herlihy, Nir Shavit, Victor Luchangco, and Michael Spear. 2021. *The Art of Multiprocessor Programming*. Morgan Kaufmann. <https://doi.org/10.1016/C2011-0-06993-4>
- [20] Maurice P. Herlihy and Jeannette M. Wing. 1990. Linearizability: A Correctness Condition for Concurrent Objects. *ACM Transactions on Programming Languages and Systems* 12, 3 (July 1990). <https://doi.org/10.1145/78969.78972>
- [21] Lorenz Hübschle-Schneider and Peter Sanders. 2016. Communication Efficient Algorithms for Top-k Selection Problems. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. <https://doi.org/10.1109/IPDPS.2016.45>
- [22] Donald B. Johnson. 1975. Priority Queues with Update and Finding Minimum Spanning Trees. *Inform. Process. Lett.* 4, 3 (Dec. 1975). [https://doi.org/10.1016/0020-0190\(75\)90001-0](https://doi.org/10.1016/0020-0190(75)90001-0)
- [23] Richard M. Karp and Yanjun Zhang. 1993. Randomized Parallel Algorithms for Backtrack Search and Branch-and-Bound Computation. *J. ACM* 40, 3 (July 1993). <https://doi.org/10.1145/174130.174145>
- [24] Jonatan Lindén and Bengt Jonsson. 2013. A Skiplist-Based Concurrent Priority Queue with Minimal Memory Contention. In *Principles of Distributed Systems (OPODIS)*. https://doi.org/10.1007/978-3-319-03850-6_15
- [25] Tobias Maier, Peter Sanders, and Roman Dementiev. 2018. Concurrent Hash Tables: Fast and General(?)! *ACM Transactions on Parallel Computing* 5, 4 (Dec. 2018). <https://doi.org/10.1145/3309206>
- [26] Silvano Martello and Paolo Toth. 1990. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Inc.
- [27] Jeff Matocha and Tracy Camp. 1998. A Taxonomy of Distributed Termination Detection Algorithms. *Journal of Systems and Software* 43, 3 (Nov. 1998). [https://doi.org/10.1016/S0164-1212\(98\)10034-1](https://doi.org/10.1016/S0164-1212(98)10034-1)

- [28] Friedemann Mattern. 1987. Algorithms for Distributed Termination Detection. *Distributed Computing* 2, 3 (Sept. 1987). <https://doi.org/10.1007/BF01782776>
- [29] Kurt Mehlhorn and Stefan Näher. 1990. Bounded Ordered Dictionaries in $O(\log \log N)$ Time and $O(n)$ Space. *Inform. Process. Lett.* 35, 4 (Aug. 1990). [https://doi.org/10.1016/0020-0190\(90\)90022-P](https://doi.org/10.1016/0020-0190(90)90022-P)
- [30] U. Meyer and P. Sanders. 2003. Δ -Stepping: A Parallelizable Shortest Path Algorithm. *Journal of Algorithms* 49, 1 (Oct. 2003). [https://doi.org/10.1016/S0196-6774\(03\)00076-2](https://doi.org/10.1016/S0196-6774(03)00076-2)
- [31] Michael Mitzenmacher, Andréa W. Richa, and Ramesh Sitaraman. 2001. The Power of Two Random Choices: A Survey of Techniques and Results. In *Handbook of Randomized Computing*. Vol. 9. https://doi.org/10.1007/978-1-4615-0013-1_9
- [32] Donald Nguyen, Andrew Lenharth, and Keshav Pingali. 2013. A Lightweight Infrastructure for Graph Analytics. In *24th ACM Symposium on Operating Systems Principles (SOSP)*. <https://doi.org/10.1145/2517349.2522739>
- [33] Patrick O’Neil, Edward Cheng, Dieter Gawlick, and Elizabeth O’Neil. 1996. The Log-Structured Merge-Tree (LSM-tree). *Acta Informatica* 33, 4 (June 1996). <https://doi.org/10.1007/s002360050048>
- [34] Anastasiia Postnikova, Nikita Koval, Giorgi Nadiradze, and Dan Alistarh. 2022. Multi-Queues Can Be State-of-the-Art Priority Schedulers. In *27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. <https://doi.org/10.1145/3503221.3508432>
- [35] William Pugh. 1990. Skip Lists: A Probabilistic Alternative to Balanced Trees. *Commun. ACM* 33, 6 (June 1990). <https://doi.org/10.1145/78973.78977>
- [36] Martin Raab and Angelika Steger. 1998. “Balls into Bins” — A Simple and Tight Analysis. In *Randomization and Approximation Techniques in Computer Science*. https://doi.org/10.1007/3-540-49543-6_13
- [37] Hamza Rihani, Peter Sanders, and Roman Dementiev. 2014. MultiQueues: Simpler, Faster, and Better Relaxed Concurrent Priority Queues. <https://doi.org/10.48550/arXiv.1411.1209>
- [38] Hamza Rihani, Peter Sanders, and Roman Dementiev. 2015. MultiQueues: Simple Relaxed Concurrent Priority Queues. In *27th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*. <https://doi.org/10.1145/2755573.2755616>
- [39] Adones Rukundo, Aras Atalar, and Philippas Tsigas. 2019. Monotonically Relaxing Concurrent Data-Structure Semantics for Increasing Performance: An Efficient 2D Design Framework. In *33rd Symposium on Distributed Computing (DISC)*. <https://doi.org/10.4230/LIPIcs.DISC.2019.31>
- [40] Adones Rukundo and Philippas Tsigas. 2021. TSLQueue: An Efficient Lock-Free Design for Priority Queues. In *Euro-Par: Parallel Processing*. https://doi.org/10.1007/978-3-030-85665-6_24
- [41] Konstantinos Sagonas and Kjell Winblad. 2017. The Contention Avoiding Concurrent Priority Queue. In *Languages and Compilers for Parallel Computing*. https://doi.org/10.1007/978-3-319-52709-3_23
- [42] Peter Sanders. 1995. Fast Priority Queues for Parallel Branch-and-Bound. In *Parallel Algorithms for Irregularly Structured Problems*. https://doi.org/10.1007/3-540-60321-2_30
- [43] Peter Sanders. 1997. *Lastverteilungsalgorithmen für parallele Tiefensuche*. Ph.D. Dissertation. <https://doi.org/10.5445/IR/997>
- [44] Peter Sanders. 1998. Randomized Priority Queues for Fast Parallel Access. *J. Parallel and Distrib. Comput.* 49, 1 (Feb. 1998). <https://doi.org/10.1006/jpdc.1998.1429>
- [45] Peter Sanders. 2001. Fast Priority Queues for Cached Memory. *ACM Journal of Experimental Algorithmics* 5 (Dec. 2001). <https://doi.org/10.1145/351827.384249>
- [46] Peter Sanders, Kurt Mehlhorn, Martin Dietzfelbinger, and Roman Dementiev. 2019. *Sequential and Parallel Algorithms and Data Structures: The Basic Toolbox*. Springer. <https://doi.org/10.1007/978-3-030-25209-0>
- [47] N. Shavit and I. Lotan. 2000. Skiplist-Based Concurrent Priority Queues. In *14th International Parallel and Distributed Processing Symposium. (IPDPS)*. <https://doi.org/10.1109/IPDPS.2000.845994>
- [48] D.R. Shier and C. Witzgall. 1981. Properties of Labeling Methods for Determining Shortest Path Trees. *J. Res. Nat. Bur. Standards* 86, 3 (1981), 317. <https://doi.org/10.6028/jres.086.013>
- [49] P. van Emde Boas, R. Kaas, and E. Zijlstra. 1976. Design and Implementation of an Efficient Priority Queue. *Mathematical systems theory* 10, 1 (Dec. 1976). <https://doi.org/10.1007/BF01683268>
- [50] Stefan Walzer and Marvin Williams. 2024. A Simple yet Exact Analysis of the MultiQueue. <https://doi.org/10.48550/arXiv.2410.08714> [cs]
- [51] J. W. J. Williams. 1964. Algorithm 232: Heapsort. *Commun. ACM* 7, 6 (1964). <https://doi.org/10.1145/512274.512284>
- [52] Marvin Williams, Peter Sanders, and Roman Dementiev. 2021. Engineering MultiQueues: Fast Relaxed Concurrent Priority Queues. In *29th European Symposium on Algorithms (ESA)*. <https://doi.org/10.4230/LIPIcs.ESA.2021.81>
- [53] Martin Wimmer, Jakob Gruber, Jesper Larsson Träff, and Philippas Tsigas. 2015. The Lock-Free k-LSM Relaxed Priority Queue. *ACM SIGPLAN Notices* 50, 8 (Jan. 2015). <https://doi.org/10.1145/2858788.2688547>
- [54] Serif Yesil, Azin Heidarshenas, Adam Morrison, and Josep Torrellas. 2019. Understanding Priority-Based Scheduling of Graph Algorithms on a Shared-Memory Platform. In *Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. <https://doi.org/10.1145/3295500.3356160>
- [55] Guozheng Zhang, Gilead Posluns, and Mark C. Jeffrey. 2024. Multi Bucket Queues: Efficient Concurrent Priority Scheduling. In *36th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*. <https://doi.org/10.1145/3626183.3659962>

A Throughput vs. Quality

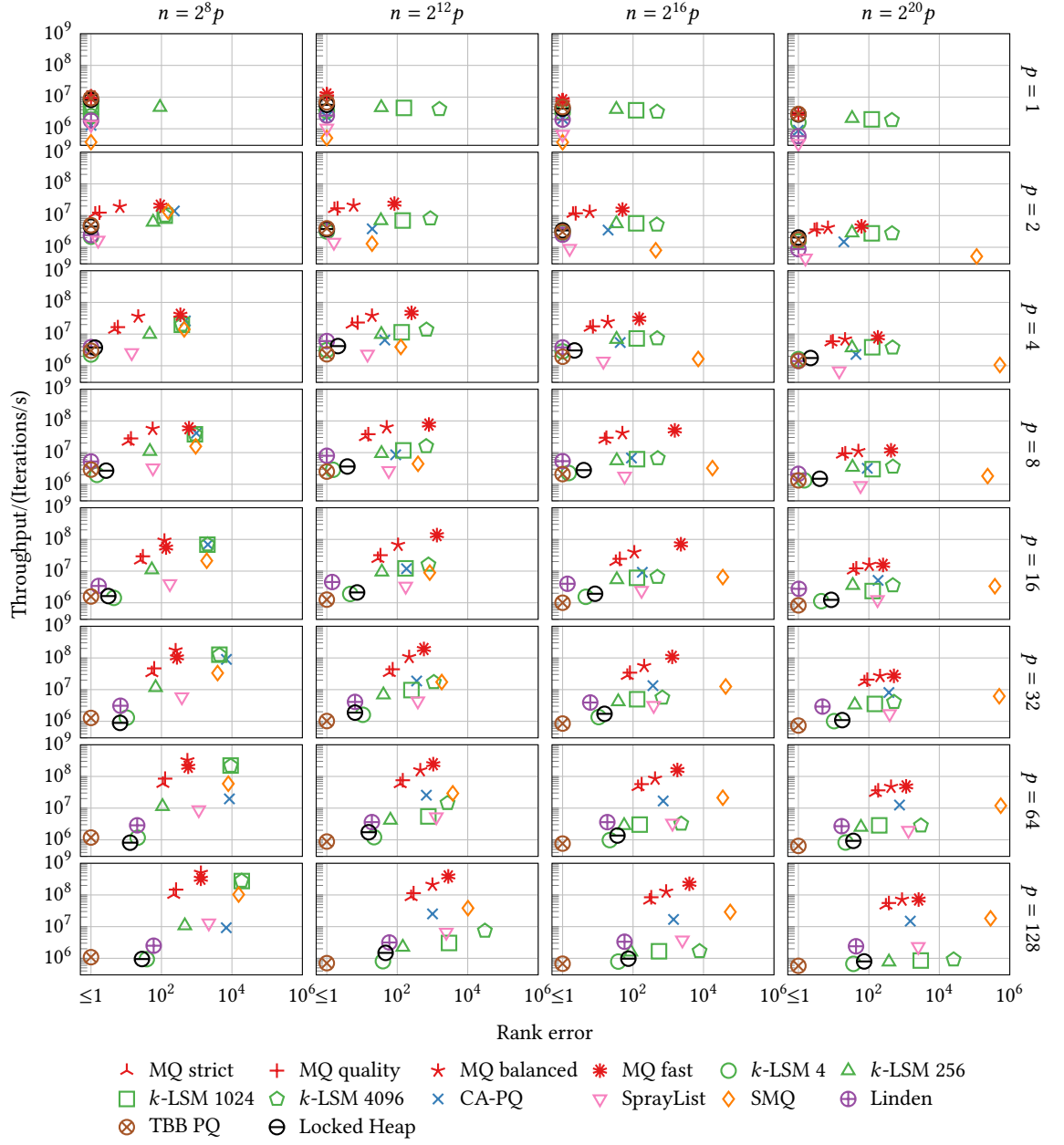


Fig. 18. Mean rank error vs. throughput for the monotonic stress test with different pre-fills n on machine AMD. Each row corresponds to a different thread count (p) from 1, 2, 4, \dots , 128 and each column to a different pre-fill. Tick labels are omitted due to space limitations.

B Throughput Scaling

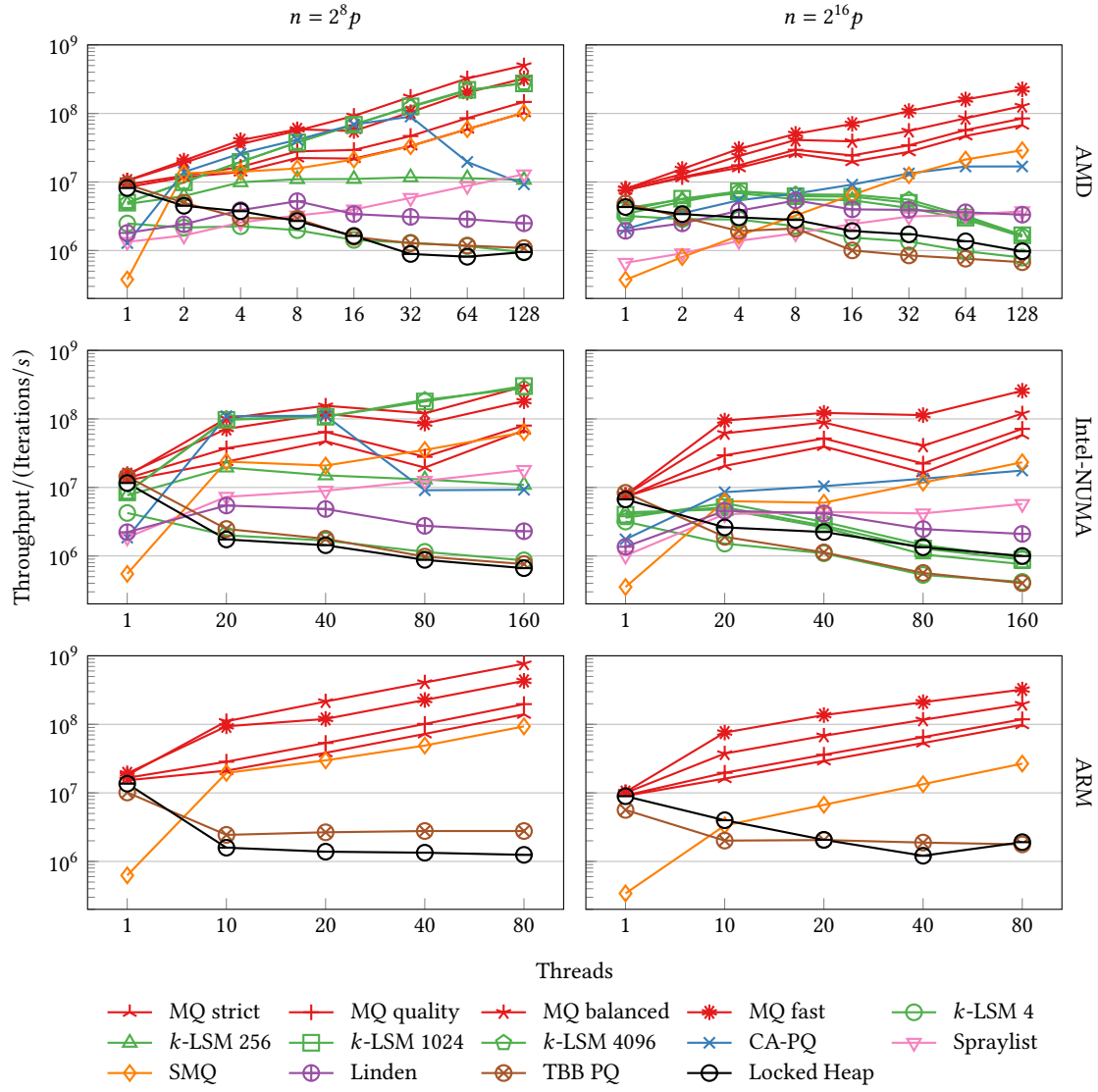


Fig. 19. Weak scaling experiment showing throughput for the monotonic stress test versus the number of threads on different machines. Each row corresponds to a different machine and each column to a different pre-fill.

C Insert-Delete Stress Test

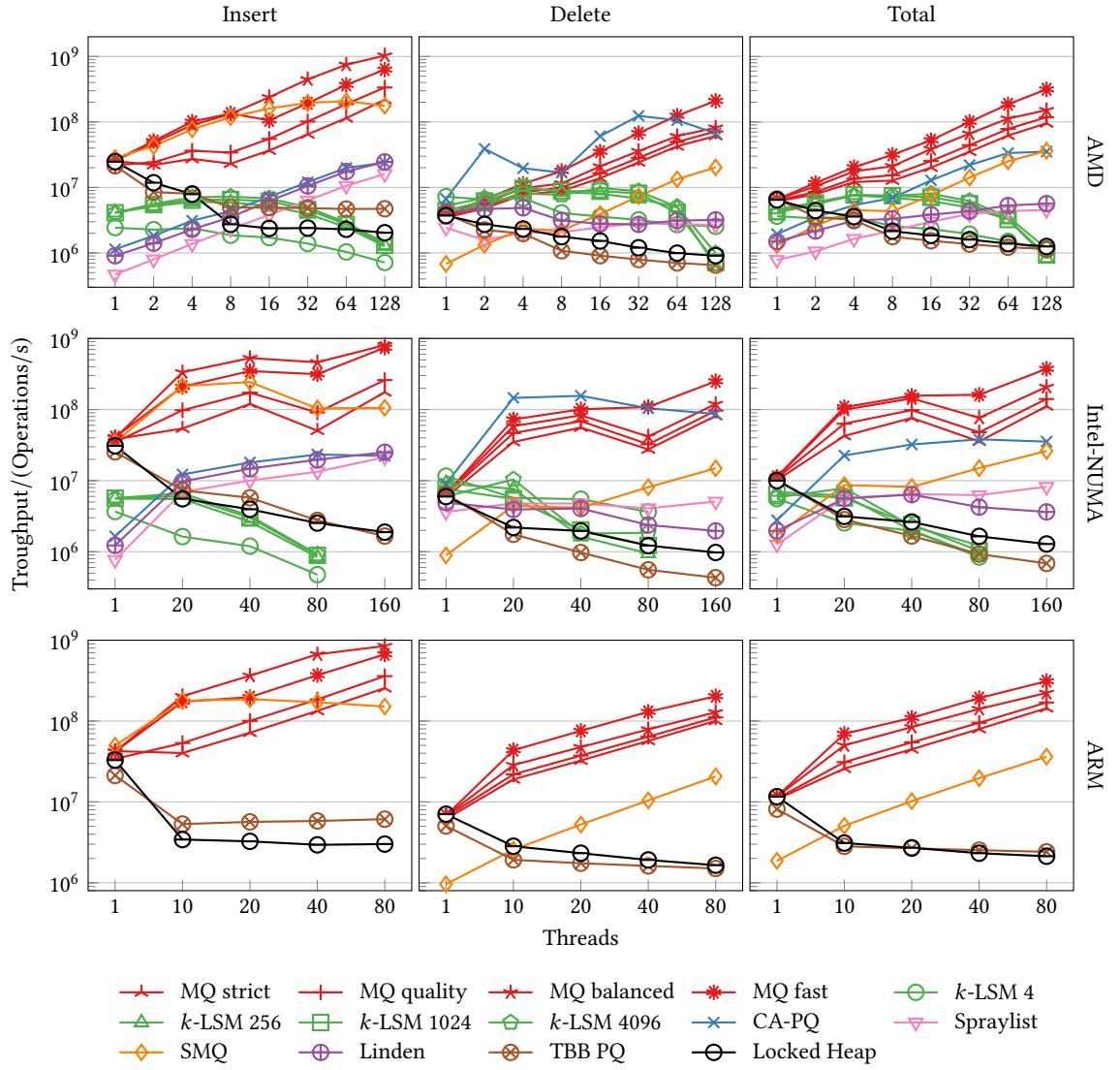


Fig. 20. Throughput comparison of the insert-delete stress test with 2^{20} elements per thread on different machines. Data points for the k -LSM variants on machine Intel-NUMA for 160 threads are missing due to crashes.

D Single-Source Shortest Path

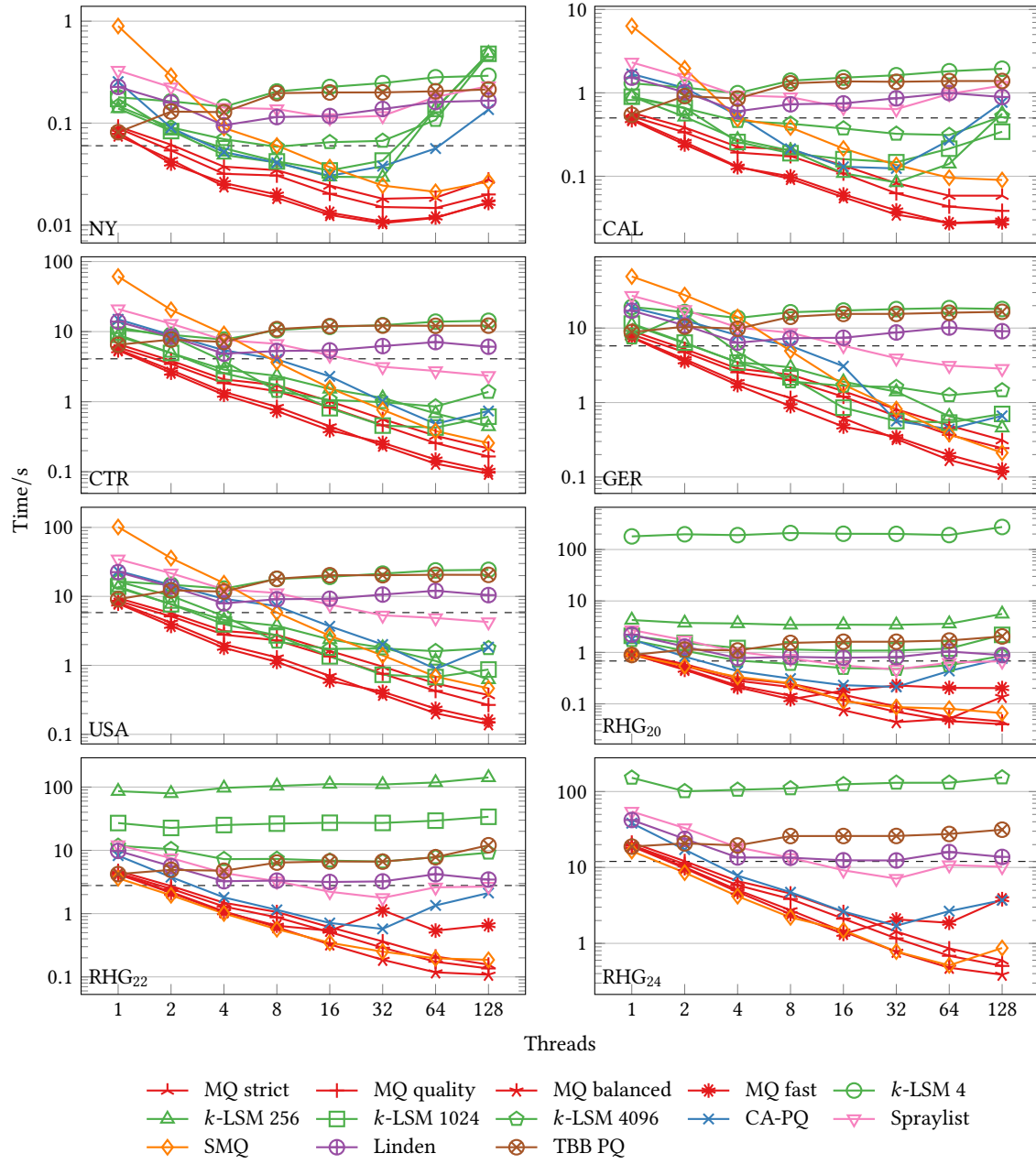


Fig. 21. Time to solve the SSSP problem with different number of threads on machine AMD. The black dashed line is the solving time of the sequential algorithm. Missing data points are due to timeouts after 300 s.

E Knapsack

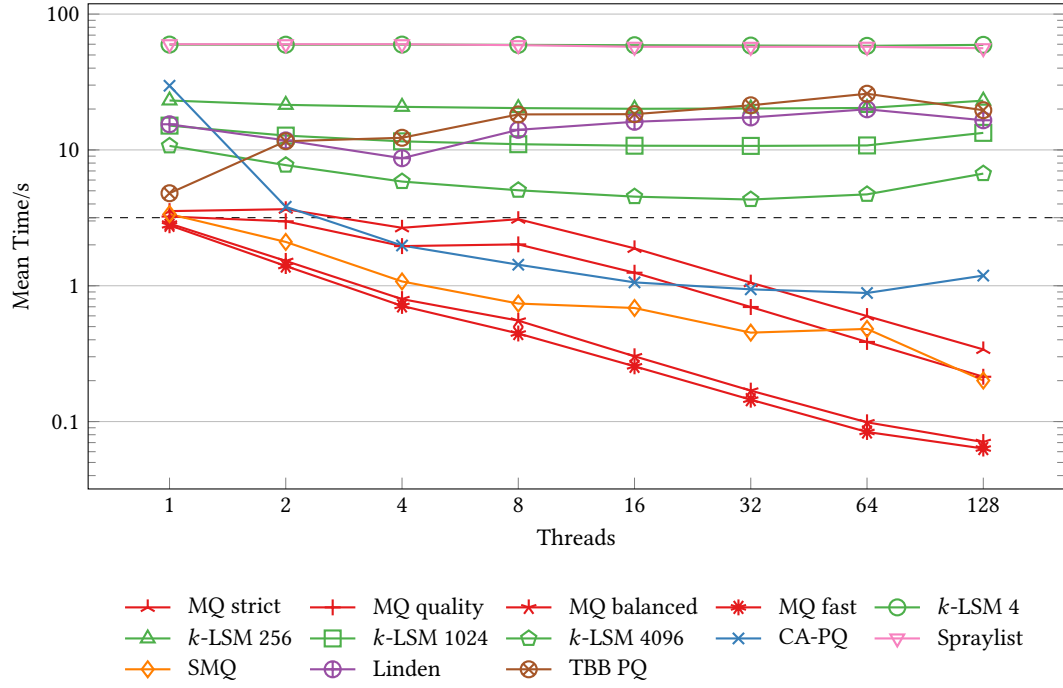


Fig. 22. Arithmetic mean of times to solve each instance of the knapsack problem with different thread counts on machine AMD. If an instance is not solved within the time limit of 60 s, the time limit is assumed. The dashed black line indicates the sequential solving time.