# The SprayList: A Scalable Relaxed Priority Queue *

Dan Alistarh
Microsoft Research

Justin Kopinsky
MIT

Jerry Li
MIT

Nir Shavit
MIT and TAU

## Abstract

High-performance concurrent priority queues are essential for applications such as task scheduling and discrete event simulation. Unfortunately, even the best performing implementations do not scale past a number of threads in the single digits. This is because of the sequential bottleneck in accessing the elements at the head of the queue in order to perform a DeleteMin operation.

In this paper, we present the `SprayList`, a scalable priority queue with relaxed ordering semantics. Starting from a non-blocking SkipList, the main innovation behind our design is that the `DeleteMin` operations avoid a sequential bottleneck by "spraying" themselves onto the head of the SkipList list in a coordinated fashion. The spraying is implemented using a carefully designed random walk, so that `DeleteMin` returns an element among the first $O(p \log^3 p)$ in the list, with high probability, where $p$ is the number of threads. We prove that the running time of a `DeleteMin` operation is $O(\log^3 p)$, with high probability, independent of the size of the list.

Our experiments show that the relaxed semantics allow the data structure to scale for high thread counts, comparable to a classic unordered SkipList. Furthermore, we observe that, for reasonably parallel workloads, the scalability benefits of relaxation considerably outweigh the additional work due to out-of-order execution.

## 1. Introduction

The necessity for increasingly efficient, scalable concurrent data structures is one of the main software trends of the past decade. Efficient concurrent implementations are known for several fundamental data structures, such as hash tables [18], linked lists [15], SkipLists [14], pools [5], and trees [6]. On the other hand, several impossibility results [3, 13] suggest that not all data structures can have efficient concurrent implementations, due to an inherent sequentiality which follows from their sequential specification.

A classic example of such a data structure is the *priority queue*, which is widely used in applications such as scheduling and event simulation, e.g. [21]. In its simplest form, a priority queue stores a set of key-value pairs, and supports two operations: `Insert`, which adds a given pair to the data structure and `DeleteMin`, which returns the key-value pair with the smallest key currently present. Sequential priority queues are well understood, and classically implemented using a heap [20]. Unfortunately, heap-based *concurrent* priority queues suffer from both memory contention and sequential bottlenecks, not only when attempting to delete the single minimal key element at the root of the heap, but also when percolating small inserted elements up the heap.

SkipList-based implementations were proposed [21, 22, 28] in order to reduce these overheads. SkipLists are randomized list-based data structures which classically support `Insert` and `Delete` operations [24]. A SkipList is composed of several linked lists organized in levels, each skipping over fewer elements. SkipLists are desirable because they allow priority queue insertions and removals without the costly percolation up a heap or the rebalancing of a search tree. Highly concurrent SkipList-based priority queues have been studied extensively and have relatively simple implementations [14, 17, 21, 25]. Unfortunately, an *exact* concurrent SkipList-based priority queue, that is, one that maintains a linearizable [17] (or even quiescently-consistent [17]) order on `DeleteMin` operations, must still remove the minimal element from the leftmost node in the SkipList. This means that all threads must repeatedly compete to decide who gets this minimal node, resulting in a bottleneck due to contention, and limited scalability [21].

An interesting alternative has been to *relax* the strong ordering constraints on the output for better scalability. An early instance of this direction is the seminal work by Karp and Zhang [19], followed up by several other interesting proposals, e.g. [7, 10, 26], designed for the (synchronous) PRAM model. Recently, there has been a surge of interest in relaxed concurrent data structures, both on the theoretical side, e.g. [16] and from practitioners, e.g. [23]. In particular, Wimmer et al. [29] explored trade-offs between ordering and scalability for asynchronous priority queues. However, despite all this effort, it is currently not clear whether it is possible to design a relaxed priority queue which provides both ordering guarantees under asynchrony, *and* scalability under high contention for realistic workloads.

In this paper, we take a step in this direction by introducing the `SprayList`, a *scalable* relaxed priority queue implementation based on a SkipList. The `SprayList` provides probabilistic guarantees on the relative priority of returned elements, and on the running time of operations. At the same time, it shows *fully scalable* throughput for up to 80 concurrent threads under high-contention workloads.

The main limitation of past SkipList-based designs was that all threads clash on the first element in the list. Instead, our idea will be to allow threads to "skip ahead" in the list, so that concurrent attempts try to remove distinct, uncontended elements. The obvious issue with this approach is that one cannot allow threads to skip ahead too far, or many high priority (minimal key) elements will not be removed.

* This paper is an extended version of work which will appear in the *Proceedings of the 20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 2015). Support is gratefully acknowledged from the National Science Foundation under grants CCF-1217921, CCF-1301926, and IIS-1447786, the Department of Energy under grant ER26116/DE-SC0008923, and the Oracle and Intel corporations.*



, .
Copyright © ACM [to be supplied]. . . $15.00.
http://dx.doi.org/10.1145/

Our solution is to have the `DeleteMin` operations traverse the SkipList, not along the list, but via a tightly controlled random walk from its root. We call this operation a *spray*. Roughly, at each SkipList level, a thread flips a random coin to decide how many nodes to skip ahead at that level. In essence, we use local randomness and the random structure of the SkipList to balance accesses to the head of the list. The lengths of jumps at each level are chosen such that the probabilities of hitting nodes among the first $O(p \log^3 p)$ are close to uniform. (See Figure 1 for the intuition behind sprays.)

While a `DeleteMin` in an exact priority queue returns the element with the smallest key—practically one of the $p$ smallest keys if $p$ threads are calling `DeleteMin` concurrently— the `SprayList` ensures that the returned key is among the $O(p \log^3 p)$ smallest keys (for some linearization of operations), and that each operation completes within $\log^3 p$ steps, both with high probability. Our design also provides anti-starvation guarantees, in particular, that elements with small keys will not remain in the queue for too long. The formal proof of these guarantees is the main technical contribution of our paper.

Specifically, our proofs are inspired by an elegant argument proving that sprays are near-uniform on an *ideal* (uniformly-spaced) SkipList, given in Section 3.2. However, this argument breaks on a realistic SkipList, whose structure is random. Precisely bounding the node hit distribution on a realistic SkipList is significantly more involved[1] Given the node hit distribution, we can upper bound the probability that two sprays collide. In turn, this upper bounds the expected number of operation retries, and the amount of time until a specific key is removed, given that a nearby key has already been removed. The uniformity of the spray distribution also allows us to implement an optimization whereby large contiguous groups of claimed nodes are physically removed by a randomly chosen *cleaner* thread.

In sum, our analysis gives strong probabilistic guarantees on the rank of a removed key, and on the running time of a `Spray` operation. Our algorithm is designed to be lock-free, but the same spraying technique would work just as well for a lock-based SkipList.

A key question is whether a priority queue with such relaxed guarantees can be useful in practice. We answer this question in the affirmative, by examining the practical performance of the SprayList through a wide series of benchmarks, including synthetic high-contention tests, discrete-event simulation, and running single-source shortest paths on grid, road network, and social graphs.

We compare our algorithm's performance to the quiescently-consistent priority queue of Lotan and Shavit [22], the state-of-the-art SkipList-based priority queue implementation of Lindén and Jonsson [21] and the recent $k$-priority queue of Wimmer et al. [29].[2]

Our first finding is that our data structure shows *fully scalable* throughput for up to 80 concurrent threads under high-contention workloads. We then focus on the trade-off between the strength of the ordering semantics and performance. We show that, for discrete-event simulation and a subset of graph workloads, the amount of additional work due to out-of-order execution is amply compensated by the increase in scalability.

**Related Work.** The first concurrent SkipList was proposed by Pugh [25], while Lotan and Shavit [22] were first to employ this data structure as a concurrent priority queue. They also noticed that the original implementation is not linearizable, and added a time-stamping mechanism for linearizability. Herlihy and Shavit [17] give a lock-free version of this algorithm.

Sundell and Tsigas [28] proposed a lock-free SkipList-based implementation which ensures linearizability by preventing threads from moving past a list element that has not been fully removed. Instead, concurrent threads help with the cleanup process. Unfortunately, all the above implementations suffer from very high contention under a standard workload, since threads are still all continuously competing for a handful of locations.

Recently, Lindén and Jonsson[21] presented an elegant design with the aim of reducing the bottleneck of deleting the minimal element. Their algorithm achieves a $30-80\%$ improvement over previous SkipList-based proposals; however, due to high contention compare-and-swap operations, its throughput does not scale past 8 concurrent threads. To the best of our knowledge, this is a limitation of all known exact priority queue implementations.

Other recent work by Mendes et al. [8] employed *elimination* techniques to adapt to contention in an effort to extend scalability. Even still, their experiments do not show throughput scaling beyond 20 threads.

Another direction by Wimmer et al. [29] presents lock-free priority queues which allow the user to dynamically decrease the strength of the ordering for improved performance. In essence, the data structure is distributed over a set of *places*, which behave as exact priority queues. Threads are free to perform operations on a place as long as the ordering guarantees are not violated. Otherwise, the thread merges the state of the place to a global task list, ensuring that the relaxation semantics hold deterministically. The paper provides analytical bounds on the work wasted by their algorithm when executing a parallel instance of Dijkstra's algorithm, and benchmark the execution time and wasted work for running parallel Dijkstra on a set of random graphs. Intuitively, the above approach provides a tighter handle on the ordering semantics than ours, at the cost of higher synchronization cost. The relative performance of the two data structures depends on the specific application scenario, and on the workload.

Our work can be seen as part of a broader research direction on high-throughput concurrent data structures with relaxed semantics [16, 27]. Examples include container data structures which (partially or entirely) forgo ordering semantics such as the rendezvous mechanism [2] or the CAFE task pool [5]. Recently, Dice et al. [11] considered randomized data structures for highly scalable exact and approximate counting.

## 2. The SprayList Algorithm

In this section, we describe the SprayList algorithm. The `Search` and `Insert` operations are identical to the standard implementations of lock-free SkipLists [14, 17], for which several freely available implementations exist, e.g. [9, 14]. In the following, we assume the reader is familiar with the structure of a SkipList, and give an overview of standard lock-free SkipList operations.

### 2.1 The Classic Lock-Free SkipList

Our presentation follows that of Fraser [14, 17], and we direct the reader to these references for a detailed presentation.

**General Structure.** The data structure maintains an implementation of a set, defined by the bottom-level lock-free list. (Throughout this paper we will use the convention that the lowest level of the SkipList is level 0.) The SkipList is comprised of multiple levels, each of which is a linked list. Every node is inserted deterministically at the lowest level, and probabilistically at higher levels. It is common for the probability that a given node is present at level $\ell$ to be $2^{-\ell}$. (Please see Figure 1 for an illustration.) A key idea in this design is that a node can be independently inserted at each level. A node is *present* if it has been inserted into the bottom list; insertion at higher levels is useful to maintain logarithmic average search time.

[1] We perform the analysis in a restricted asynchronous model, defined in Section 3.1.

[2] Due to the complexity of the framework of [29], we only provide a partial comparison with our algorithm in terms of performance.

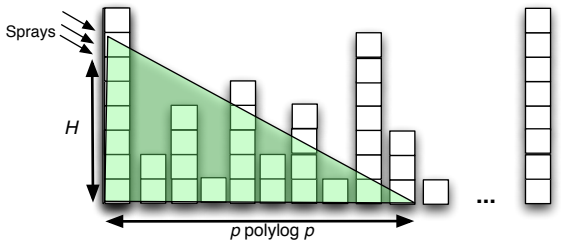


Figure 1: The intuition behind the `SprayList`. Threads start at height $H$ and perform a random walk on nodes at the start of the list, attempting to acquire the node they land on.

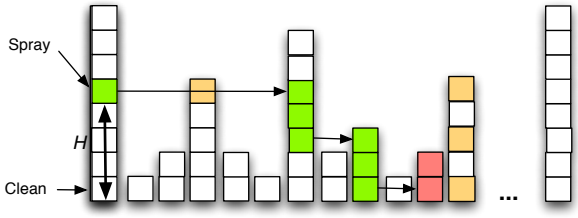


Figure 2: A simple example of a spray, with no padding. Green nodes are touched by the `Spray`, and the thread stops at the red node. Orange nodes could have been chosen for jumps, but were not.

**Pointer Marking.** A critical issue when implementing lock-free lists is that nodes might "vanish" (i.e., be removed concurrently) while some thread is trying access them. Fraser and Harris [14] solve this problem by reserving a *marked* bit in each pointer field of the SkipList. A node with a marked bit is itself *marked*. The bit is always checked and masked off before accessing the node.

**Search.** As in the sequential implementation, the SkipList search procedure looks for a *left* and *right* node at each level in the list. These nodes are adjacent om the list, with key values less-than and greater-than-equal-to the search key, respectively.

The search loop skips over marked nodes, since they have been logically removed from the list. The search procedure also helps clean up marked nodes from the list: if the thread encounters a sequence of marked nodes, these are removed by updating the unmarked successor to point to the unmarked predecessor in the list at this level. If the currently accessed node becomes marked during the traversal, the entire search is re-started from the SkipList head. The operation returns the node with the required key, if found at some level of the list, as well as the list of successors of the node.

**Delete.** Deletion of a node with key $k$ begins by first searching for the node. If the node is found, then it is *logically deleted* by updating its value field to `NULL`. The next stage is to mark each link pointer in the node. This will prevent an new nodes from being inserted after the deleted node. Finally, all references to the deleted node are removed. Interestingly, Fraser showed that this can be done by performing a *search* for the key: recall that the search procedure swings list pointers over marked nodes.

**Cleaners / Lotan-Shavit DeleteMin.** In this context, the Lotan-Shavit [22] `DeleteMin` operation traverses the bottom list attempting to acquire a node via a locking operation. Once acquired, the node is logically deleted and then removed via a search operation. We note that this is exactly the same procedure as the periodic *cleaner* operations in our design.

**Insert.** A new node is created with a randomly chosen height. The node's pointers are unmarked, and the set of successors is set to the successors returned by the *search* method on the node's key. Next, the node is inserted into the lists by linking it between the successors and the predecessors obtained by searching. The updates are performed using compare-and-swap. If a compare-and-swap fails, the list must have changed, and the call is restarted. The insert then progressively links the node up to higher levels. Once all levels are linked, the method returns.

### 2.2 Spraying and Deletion

The goal of the `Spray` operation is to emulate a uniform choice among the $O(p \log^3 p)$ highest-priority items. To perform a `Spray`, a process starts at the front of the SkipList, and at some initial height $h$. (See Figure 2 for an illustration.)

At each horizontal level $\ell$ of the list, the process first jumps forward for some small, randomly chosen number of steps $j_\ell \geq 0$. After traversing those nodes, the process descends some number of levels $d_\ell$, then resumes the horizontal jumps. We iterate this procedure until the process reaches a node at the bottom of the SkipList.

Once on the bottom list, the process attempts to acquire the current node. If the node is successfully acquired, the thread starts the standard SkipList removal procedure, marking the node as logically deleted. (As in the SkipList algorithm, logically deleted nodes are ignored by future traversals.) Otherwise, if the process fails to acquire the node, it either re-tries a `Spray`, or, with low probability, becomes a *cleaner* thread, searching linearly through the bottom list for an available node.

We note that, as with other SkipList based Priority Queue algorithms, the runtime of a `Spray` operation is independent of the size of the SkipList. This is because, with high probability, the `Spray` operation only accesses pointers belonging to the $O(p \log^3 p)$ items at the head of the list.

**Spray Parameters.** An efficient `Spray` needs the right combination of parameters. In particular, notice that we can vary the starting height, the distribution for jump lengths at each level, and how many levels to descend between jumps. The constraints are poly-logarithmic time for a `Spray`, and a roughly uniform distribution over the head of the list. At the same time, we need to balance the average length of a `Spray` with the expected number of thread collisions on elements in the bottom list.

We now give an overview of the parameter choices for our implementation. For simplicity, consider a SkipList on which no removes have yet occurred due to `Spray` operations. We assume that the data structure contains $n$ elements, where $n \gg p$.

**Starting Height.** Each `Spray` starts at list level $H = \log p + K$, for some constant $K$.[3] (Intuitively, starting the `Spray` from a height less than $\log p$ leads to a high number of collisions, while starting from a height of $C \log p$ for $C > 1$ leads to `Spray`s which traverse beyond the first $O(p \log^3 p)$ elements.)

**Jump Length Distribution.** We choose the maximum number of forward steps $L$ that a `Spray` may take at a level to be $L = M \log^3 p$, where $M \geq 1$ is a constant. Thus, the number of forward steps at level $\ell$, is uniformly distributed in the interval $[0, L]$.

The intuitive reason for this choice is that a randomly built SkipList is likely to have chains of $\log p$ consecutive elements of height one, which can only be accessed through the bottom list. We wish to be able to choose uniformly among such elements, and we therefore need $L$ to be at least $\log p$. (While the same argument does not apply at higher levels, our analysis shows that choosing this jump length $j_\ell$ yields good uniformity properties.)

**Levels to Descend.** The final parameter is the choice of how many levels to descend after a jump. A natural choice, used in

[3] Thoughout this paper, unless otherwise stated, we consider all logarithms to be integer, and omit the floor $\lfloor \cdot \rfloor$ notation.

our implementation, is to descend one level at a time, i.e., perform horizontal jumps at each SkipList level.

In the analysis, we consider a slightly more involved random walk, which descends $D = \max(1, \lfloor \log \log p \rfloor)$ consecutive levels after a jump at level $\ell$. We must always traverse the bottom level of the SkipList (or we will never hit SkipList nodes of height 1) so we round $H$ down to the nearest multiple of $D$. We note that we found empirically that setting $D = 1$ yields similar performance.

In the following, we parametrize the implementation by $H$, $L$ and $D$ such that $D$ evenly divides $H$. The pseudocode for $\texttt{Spray}(H, L, D)$ is given below.

```
x ← head          /* x = pointer to current location */
                          /* Assume D divides H */
ℓ ← H                   /* ℓ is the current level  */
while ℓ ≥ 0 do
    Choose j_ℓ ← Uniform[0, L]          /* random jump */
    Walk x forward j_ℓ steps on list at height ℓ
                  /* traverse the list at this level */
    ℓ ← ℓ − D                    /* descend D levels */
Return x
```

**Algorithm 1:** Pseudocode for $\texttt{Spray}(H, L, D)$. Recall that the bottom level of the SkipList has height 0.

**Node Removal.** Once it has successfully acquired a node, the thread proceeds to remove it as in a standard lock-free SkipList [14, 17]. More precisely, the node is logically deleted, and its references are marked as invalid.

In a standard implementation, the final step would be to swing the pointers from its predecessor nodes to its successors. However, a spraying thread skips this step and returns the node. Instead, the pointers will be corrected by *cleaner* threads: these are randomly chosen `DeleteMin` operations which linearly traverse the bottom of the list in order to find a free node, as described in Section 2.3.

### 2.3 Optimizations

**Padding.** A first practical observation is that, for small (constant) values of $D$, the `Spray` procedure above is biased against elements at the front of the list. For example, it would be extremely unlikely for the second element in the list to be hit by a walk for $D = 1$. To circumvent this bias, in such cases, we simply "pad" the SkipList: we add $K(p)$ dummy entries in the front of the SkipList. If a `Spray` lands on one of the first $K(p)$ dummy entries, it restarts. We choose $K(p)$ such that the restart probability is low, while, at the same time, the probability that a node in the interval $[K(p) + 1, p \log^3 p]$ is hit is close to $1/p \log^3 p$. We note that padding is not necessary for higher values of $D$, e.g., $D = \Theta(\log \log p)$.

**Cleaners.** Before each new `Spray`, each thread flips a low-probability coin to decide whether it will become a *cleaner* thread. A cleaner thread simply traverses the bottom-level list of the SkipList linearly (skipping the padding nodes), searching for a key to acquire. In other words, a cleaner simply executes a lock-free version of the Lotan-Shavit [22] `DeleteMin` operation. At the same time, notice that cleaner threads adjust pointers for nodes previously acquired by other `Spray` operations, reducing contention and wasted work. Interestingly, we notice that a cleaner thread can swing pointers across a whole group of nodes that have been marked as logically deleted, effectively batching this part of the remove process.

The existence of cleaners is not needed in the analysis, but is a useful optimization. In the implementation, the probability of an operation becoming a cleaner is $1/p$, i.e., roughly one in $p$ `Spray`s becomes a cleaner.

**Adapting to Contention.** We also note that the `SprayList` allows threads to adjust the spray parameters based on the level of contention. In particular, a thread can estimate $p$, increasing its estimate if it detects higher than expected contention (in the form of collisions) and decreasing its estimate if it detects low contention. Each thread parametrizes its `Spray` parameters the same way as in the static case, but using its estimate of $p$ rather than a known value. Note that with this optimization enabled, if only a single thread access the `SprayList`, it will always dequeue the element with the smallest key.

## 3. Spray Analysis

In this section, we analyze the behavior of `Spray` operations. We describe our analytical model in Section 3.1. We then give a first motivating result in Section 3.2, bounding the probability that two `Spray` operations collide for an ideal SkipList.

We state our main technical result, Theorem 3, and provide a proof overview in Section 3.3. The full proof of Theorem 3 is rather technical, and can be found in the non-anonymous supplemental material submitted with this paper. In essence, given our model, our results show that SprayLists do not return low priority elements except with extremely small probability (Theorem 2) and that there is very low contention on individual elements, which in turn implies the bound on the running time of `Spray` (Corollary 1).

### 3.1 Analytical Model

As with other complex concurrent data structures, a complete analysis of spraying in a fully asynchronous setting is extremely challenging. Instead, we restrict our attention to showing that, under reasonable assumptions, spraying approximates uniform choice amongst roughly the first $O(p \log^3 p)$ elements. We will then use this fact to bound the contention between `Spray` operations. We therefore assume that there are $n \gg p \log^3 p$ elements in the SkipList.

We consider a set of at most $p$ concurrent, asynchronous processes trying to perform `DeleteMin` operations, traversing a *clean* SkipList, i.e. a SkipList whose height distribution is the same as one that has just been built. In particular, a node has height $\geq i$ with probability $1/2^i$, independent of all other nodes. They do so by each performing `Spray` operations. When two or more `Spray` operations end at the same node, all but one of them must retry. if a `Spray` lands in the padded region of the SkipList, it must also retry. We repeat this until all `Spray`s land at unique nodes (because at most one thread can obtain a node). Our goal is to show that for all $p$ processors, this process will terminate in $O(\log^3 p)$ time in expectation and with high probability. Note that since each `Spray` operation takes $O(\log^3 p)$ time, this is equivalent to saying that each process must restart their `Spray` operations at most a constant number of times, in expectation and with high probability. We guarantee this by showing that `Spray` operations have low contention.

On the one hand, this setup is clearly only an approximation of a real execution, since concurrent inserts and removes may occur in the prefix and change the SkipList structure. Also, the structure of the list may have been biased by previous `Spray` operations. (For example, previous sprays might have been biased to land on nodes of large height, and therefore such elements may be less probable in a dynamic execution.)

On the other hand, we believe this to be a reasonable approximation for our purposes. We are interested mainly in spray distribution; concurrent deletions should not have a high impact, since, by the structure of the algorithm, logically deleted nodes are skipped by the spray. Also, in many scenarios, a majority of the concurrent inserts are performed towards the back of the list (corresponding to elements of lower priority than those at the front). Finally, the effect of the spray distribution on the height should be limited, since removing an element uniformly at random from the list does not change its expected structure, and we closely approximate uniform removal. Also, notice that cleaner threads (linearly traversing

the bottom list) periodically "refresh" the SkipList back to a clean state.

### 3.2 Motivating Result: Analysis on a Perfect SkipList

In this section, we illustrate some of the main ideas behind our runtime argument by first proving a simpler claim, Theorem 1, which holds for an idealized SkipList. Basically, Theorem 1 says that, on SkipList where nodes of the same height are evenly spaced, the `Spray` procedure ensures low contention on individual list nodes.

More precisely, we say a SkipList is *perfect* if the distance between any two consecutive elements of height $\geq j$ is $2^j$, and the first element has height 0. On a perfect SkipList, we do not have to worry about probability concentration bounds when considering SkipList structure, which simplifies the argument. (We shall take these technicalities into account in the complete argument in the next section.)

We consider the $\texttt{Spray}(H, L, D)$ procedure with parameters $H = \log p - 1$, $L = \log p$, and $D = 1$, the same as our implementation version. Practically, the walk starts at level $\log p - 1$ of the SkipList, and, at each level, uniformly chooses a number of forward steps between $[1, \log p]$ before descending. We prove the following upper bound on the collision probability, assuming that $\log p$ is even:

**Theorem 1.** *For any position $x$ in a perfect SkipList, let $F_p(x)$ denote the probability that a $\texttt{Spray}(\log p - 1, \log p, 1)$ lands at $x$. Then*

$$F_p(x) \leq 1/(2p).$$

*Proof.* In the following, fix parameters $H = \log p - 1, L = \log p, D = 1$ for the `Spray`, and consider an arbitrary such operation. Let $a_i$ be the number of forward steps taken by the `Spray` at level $i$, for all $0 \leq i \leq \log p - 1$.

We start from the observation that, on a *perfect* SkipList, the operation lands at the element of index $\sum_{i=0}^{\log p - 1} a_i 2^i$ in the bottom list. Thus, for any element index $x$, to count the probability that a `Spray` which lands at $x$, it suffices to compute the probability that a $(\log p + 1)$-tuple $(a_0, \ldots, a_{\log p})$ whose elements are chosen independently and uniformly from the interval $\{1, \ldots, \log p\}$ has the property that the jumps sum up to $x$, that is,

$$\sum_{i=0}^{\log p - 1} a_i 2^i = x. \tag{1}$$

For each $i$, let $a_i(j)$ denote the $j$th least significant bit of $a_i$ in the binary expansion of $a_i$, and let $x(j)$ denote the $j$th least significant bit of $x$ in its binary expansion.

Choosing an arbitrary `Spray` is equivalent to choosing a random $(\log p)$-tuple $(a_1, \ldots, a_{\log p})$ as specified above. We wish to compute the probability that the random tuple satisfies Equation 1. Notice that, for $\sum_{i=0}^{\log p - 1} a_i 2^i = x$, we must have that $a_0(1) = x(1)$, since the other $a_i$ are all multiplied by some nontrivial power of 2 in the sum and thus their contribution to the ones digit (in binary) of the sum is always 0. Similarly, since all the $a_i$ except $a_0$ and $a_1$ are bit-shifted left at least twice, this implies that if Equation 1 is satisfied, then we must have $a_1(1) + a_0(2) = x(2)$. In general, for all $1 \leq k \leq \log p - 1$, we see that to satisfy Equation 1, we must have that $a_k(1) + a_{k-1}(2) + \ldots + a_0(k) + c = x(k)$, where $c$ is a carry bit determined completely by the choice of $a_0, \ldots, a_{k-1}$.

Consider the following random process: in the 0th round, generate $a_0$ uniformly at random from the interval $\{1, \ldots, \log p\}$, and test if $a_0(1) = x(1)$. If it satisfies this condition, continue and we say it passes the first round, otherwise, we say we fail this round. Iteratively, in the $k$th round, for all $1 \leq k \leq \log p - 1$, randomly generate an $a_k$ uniformly from the interval $\{1, \ldots, \log p\}$, and check that $a_k(1) + a_{k-1}(2) + \ldots + a_0(k) + c = x(k) \mod 2$, where $c$ is the carry bit determined completely by the choice of $a_0, \ldots, a_{k-1}$ as described above. If it passes this test, we continue and say that it passes the $k$th round; otherwise, we fail this round. If we have yet to fail after the $(\log p - 1)$st round, then we output `PASS`, otherwise, we output `FAIL`. By the argument above, the probability that we output `PASS` with this process is an upper bound on the probability that a `Spray` lands at $x$.

The probability we output `PASS` is then

$$\Pr[\text{pass 0th round}] \prod_{i=0}^{\log p - 2} \Pr[\text{pass } (i+1)\text{th round} | A_i]$$

where $A_i$ is the event that we pass all rounds $k \leq i$. Since $a_0$ is generated uniformly from the interval $\{1, 2, \ldots, \log p\}$, and since $\log p$ is even by assumption, the probability that the least significant bit of $a_0$ is $x(1)$ is exactly $1/2$, so

$$\Pr[\text{pass 0th round}] = 1/2. \tag{2}$$

Moreover, for any $1 \leq i \leq \log p - 2$, notice that conditioned on the choice of $a_1, \ldots, a_i$, the probability that we pass the $(i+1)$th round is exactly the probability that the least significant bit of $a_{i+1}$ is equal to $x(i+1) - (a_i(2) + \ldots + a_0(i+1) + c) \mod 2$, where $c$ is some carry bit as we described above which only depends on $a_1, \ldots, a_i$. But this is just some value $v \in \{0, 1\}$ wholly determined by the choice of $a_0, \ldots, a_i$, and thus, conditioned on any choice of $a_0, \ldots, a_i$, the probability that we pass the $(i+1)$th round is exactly $1/2$ just as above. Since the condition that we pass the $k$th round for all $k \leq i$ only depends on the choice of$a_0, \ldots, a_i$, we conclude that

$$\Pr[\text{pass } (i+1)\text{th round} | A_i] = 1/2. \tag{3}$$

Therefore, we have $\Pr[\text{output } \texttt{PASS}] = (1/2)^{\log p} = 1/p$, which completes the proof.

□

### 3.3 Complete Runtime Analysis for `DeleteMin`

In this section, we show that, given a randomly chosen SkipList, each `DeleteMin` operation completes in $O(\log^3 p)$ steps, in expectation and with high probability. As mentioned previously, this is equivalent to saying that the `Spray` operations for each process restart at most a constant number of times, in expectation and with high probability. The crux of this result (stated in Corollary 1) is a characterization of the probability distribution induced by `Spray` operations on an arbitrary SkipList, which we obtain in Theorem 3. Our results require some mathematical preliminaries. For simplicity of exposition, throughout this section and in the full analysis we assume $p$ which is a power of 2. (If $p$ is not a power of two we can instead run `Spray` with the $p$ set to the smallest power of two larger than the true $p$, and incur a constant factor loss in the strength of our results.)

We consider `Spray`s with the parameters $H = \log p - 1$, $L = M \log^3 p$, and $D = \max(1, \log \log p)$. We will assume that all jump parameters are integers, and that $D$ divides $H$. The claim is true even when these assumptions do not hold, but we only present the analysis in this special case because the presentation otherwise becomes too messy. Let $\ell_p$ be the number of levels at which traversals are performed, except the bottom level; in particular $\ell_p = H/D$.

Since we only care about the relative ordering of the elements in the SkipList with each other and not their real priorities, we will call the element with the $i$th lowest priority in the SkipList the $i$th element in the SkipList. We will also need the following definition.

**Definition 1.** *Fix two positive functions $f(p), g(p)$.*

- *We say that $f$ and $g$ are* asymptotically equal*, $f \simeq g$, if* $\lim_{p \to \infty} f(p)/g(p) = 1$.

- *We say that $f \lesssim g$, or that $g$* asymptotically bounds *$f$, if there exists a function $h \simeq 1$ so that $f(p) \leq h(p)g(p)$ for all $p$.*

Note that saying that $f \simeq g$ is stronger than saying that $f = \Theta(g)$, as it insists that the constant that the big-Theta would hide is in fact 1, i.e. that asymptotically, the two functions behave exactly alike even up to constant factors.

There are two sources of randomness in the `Spray` algorithm and thus in the statement of our theorem. First, there is the randomness over the choice of the SkipList. Given the elements in the SkipList, the randomness in the SkipList is over the heights of the nodes in the SkipList. To model this rigorously, for any such SkipList $S$, we identify it with the $n$-length vectors $(h_1, \ldots, h_n)$ of natural numbers (recall there are $n$ elements in the SkipList), where $h_i$ denotes the height of the $i$th node in the SkipList. Given this representation, the probability that $S$ occurs is $\prod_{i=1}^{n} 2^{-(h_i)}$.

Second, there is the randomness of the `Spray` algorithm itself. Formally, we identify each `Spray` with the $(\ell_p + 1)$-length vector $(a_0, \ldots, a_{\ell_p})$ where $1 \leq a_i \leq M \log^3 p$ denotes how far we walk at height $iD$, and $a_0$ denotes how far we walk at the bottom height. Our `Spray` algorithm uniformly chooses a combination from the space of all possible `Spray`s. For a fixed SkipList $S$, and given a choice for the steps at each level in the `Spray`, we say that the `Spray` *returns* element $i$ if, after doing the walk prescribed by the lengths chosen and the procedure described in Algorithm 1, we end at element $i$. For a fixed SkipList $S \in \mathcal{S}$ and some element $i$ in the SkipList, we let $F_p(i, S)$ denote the probability that a `Spray` returns element $i$. We will write this often as $F_p(i)$ when it is clear which $S$ we are working with.

**Definition 2.** *We say an event happens* with high probability *or* w.h.p. *for short if it occurs with probability at least $1 - p^{-\Omega(M)}$, where $M$ is the constant defined in Algorithm 1.*

#### 3.3.1 Top Level Theorems

With these definitions we are now equipped to state our main theorems about SprayLists.

**Theorem 2.** *In the model described above, no* `Spray` *will return an element beyond the first $M(1 + \frac{1}{\log p})\sigma(p)p \log^3 p \simeq Mp \log^3 p$, with probability at least $1 - p^{-\Omega(M)}$.*

This theorem states simply that sprays do not go too far past the first $O(p \log^3 p)$ elements in the SkipList, which demonstrates that our SprayList does return elements with relatively small priority. The proof of Theorem 2 is fairly straightforward and uses standard concentration bounds and is available in the non-anonymous supplemental material submitted with this paper. However, the tools we use there will be crucial to later proofs. The other main technical contribution of this paper is the following theorem.

**Theorem 3.** *For $p \geq 2$ and under the stated assumptions, there exists an interval of elements $I(p) = [a(p), b(p)]$ of length $b(p) - a(p) \simeq Mp \log^3 p$ and endpoint $b(p) \lesssim Mp \log^3 p$, such that for all elements in the SkipList in the interval $I(p)$, we have that*

$$F_p(i, S) \simeq \frac{1}{Mp \log^3 p},$$

*w.h.p. over the choice of $S$.*

In plain words, this theorem states that there exists a range of elements $I(p)$, whose length is asymptotically equal to $Mp \log^3 p$, such that if you take a random SkipList, then with high probability over the choice of that SkipList, the random process of performing `Spray` approximates uniformly random selection of elements in the range $I(p)$, up to a factor of two. The condition $b(p) \lesssim Mp \log^3 p$ simply means that the right endpoint of the interval is not very far to the right. In particular, if we pad the start of the SkipList with $K(p) = a(p)$ dummy elements, the `Spray` procedure will approximate uniform selection from roughly the first $Mp \log^3 p$ elements, w.h.p. over the random choice of the SkipList. The proof of Theorem 3 is takes up .

**Runtime Bound.** Given this theorem, we then use it to bound the probability of collision for two Sprays, which in turn bounds the running time for a `DeleteMin` operation, which yields the following Corollary. Given Theorem 3, its proof is fairly straightforward. We give its proof in Section 3.3.6.

**Corollary 1.** *In the model described above,* `DeleteMin` *takes $O(\log^3 p)$ time in expectation. Moreover, for any $\epsilon > 0$,* `DeleteMin` *will run in time $O(\log^3 p \log \frac{1}{\epsilon})$ with probability at least $1 - \epsilon$.*

#### 3.3.2 Proof of Theorem 2

Throughout the rest of the section, we will need a way to talk about partial `Spray`s, those which have only completely some number of levels.

**Definition 3.** *Fix a `Spray` $S$, $(a_0, \ldots, a_{\ell_p})$ where $1 \leq a_i \leq M \log^3 p$.*

- *To any $k$-tuple $(b_k, \ldots, b_\ell)$ for $k \geq 0$, associate to it the walk which occurs if, descending from level $\ell_p D$, we take $b_r$ steps at each height $rD$, as specified in `Spray`. We define the $k$-prefix of $S$ to be the walk associated with $(a_k, \ldots, a_\ell)$. We say the $k$-prefix of $S$* returns *the element that the walk described ends at.*
- *To any $(k+1)$-tuple $(b_0, \ldots, b_k)$ for $k \leq \ell_p$ and any starting element $i$, associate to it the walk which occurs if, descending from level $kD$, we take $b_r$ steps at each height $rD$, as specified in `Spray`. We define the $k$-suffix of $S$ to be the walk associated with $(a_0, \ldots, a_k)$, starting at the node the $(\ell - k - 1)$-prefix of $S$ returns. We say the $k$-prefix of $S$* returns *the element that the walk described ends at.*
- *The $k$th* part *of $S$ is the walk at level $kD$ of length $a_k$ starting at the element that the $(\ell_p - k + 1)$-prefix of $S$ returns.*

Intuitively, the $k$-prefix of a spray is simply the walk performed at the $k$ top levels of the spray, and the $k$-suffix of a spray is simply the walk at the bottom $k$ levels of the spray.

For $k \geq 0$, let $E_k$ denote the expected distance the `Spray` travels at the $kD$th level if it jumps exactly $M \log^3 p$ steps. In particular,

$$E_k = M 2^{kD} \log^3 p.$$

We in fact prove the following, stronger version of Theorem 2.

**Lemma 1.** *Let $\sigma(p) = \log p/(\log p - 1)$. For any fixed $\alpha$, the $k$-suffix of any `Spray` will go a distance of at most $(1 + \alpha)\sigma(p)E_{kD+1}$, with probability at least $1 - p^{-\Omega(M\alpha^2 \log^2 p)}$ over the choice of the SkipList. To prove this we first need the following proposition.*

Notice that setting $\alpha = 1/\log p$ and $k = \ell_p$ then gives us the Theorem 2. Thus it suffices to prove Lemma 1. First, we require a technical proposition.

**Proposition 1.** *For $k \leq \log p$ and $\alpha > 0$, the probability that the $k$th part of a* `Spray` *travels more than $(1+\alpha)M2^k \log^3 p$ distance is at most $(1/p)^{\Omega(M\alpha^2 \log^2 p)}$.*

*Proof.* Fix some node $x$ in the list. Let $X_T$ be the number of elements with height at least $k$ that we encounter in a random walk of $T$ steps starting at $x$. We know that $\mathbb{E}(X_T) = T/2^k$. Choose $T = (1 + \alpha)M2^k \log^3 p$. Then by a Chernoff bound, $\Pr(X_T \leq (1 + \alpha)M \log^3 p) \leq p^{-\Omega(M\alpha^2 \log^2 p)}$.

Therefore, if we take $T$ steps at the bottom level we will with high probability hit enough elements of the necessary height, which

implies that a Spray at that height will not go more than that distance. □

*Proof of Lemma 1.* WLOG suppose we start start at the head of the list, and $j$ is the element with distance $(1+\alpha)\sigma(p)E_{kD+1}$ from the head. Consider the hypothetical Spray which takes the maximal number of allowed steps at each level $rD$ for $r \leq k$. Clearly this Spray goes the farthest of any Spray walking at levels $kD$ and below, so if this Spray cannot reach $j$ starting at the front of the list and walking only on levels $kD$ and below, then no Spray can. Let $x_r$ denote the element at which the Spray ends up after it finishes its $rD$th level for $0 \leq r \leq k$ and $x_{kD+1} = 0$, and let $d_r$ be the distance that the Spray travels at level $rD$. For any $r \geq 0$, by Proposition 1 and the union bound, $\Pr(\exists k : d_r > (1+\alpha)E_{rD}) \leq p^{-\Omega(M\alpha^2 \log^2 p)}$.

Therefore, w.h.p., the distance that this worst-case spray will travel is upper bounded by

$$\begin{aligned}\sum_{r=0}^{k} d_r &\leq (1+\alpha)\sum_{r=0}^{k} E_r \\ &\leq (1+\alpha)\sigma(p)E_{kD+1}.\end{aligned}$$

□

#### 3.3.3 Outline of Proof of Theorem 3

We prove Theorem 3 by proving the following two results:

**Lemma 2.** *For all elements $i$, we have*

$$F_i(p,S) \lesssim \frac{1}{pM\log^3 p}$$

*with high probability over the choice of $S$.*

**Lemma 3.** *There is some constant $A > 1$ which for $p$ sufficiently large can be chosen arbitrarily close to 1 so that for all*

$$i \in [Ap\log^2 p, \frac{1}{1+1/\log p} Mp\log^3 p],$$

*we have*

$$F_i(p) \gtrsim \frac{1}{Mp\log^3 p}$$

*with high probability over the choice of $S$.*

Given these two lemmas, if we then let $I(p)$ be the interval defined in Theorem 3, it is straightforward to argue that this interval satisfies the desired properties for Theorem 3 w.h.p. over the choice of the SkipList $S$. Thus the rest of this section is dedicated to the proofs of these two lemmas.

Fix any interval $I = [a,b]$ for $a, b \in \mathbb{N}$ and $a \leq b$. In expectation, there are $(b-a+1)2^{k-1}$ elements in $I$ with height at least $k$ in the SkipList; the following Lemma bounds the deviation from the expectation.

**Proposition 2.** *For any $b$, and any height $h$, let $D_{b,h}$ be the number of items between the $(b-k)$th item and the $b$th item in the SkipList with height at least $h$, and let $E_{b,h} = (k+1)2^{1-h}$ be the expected value of $D_{b,h}$. Then for any $\alpha > 0$,*

$$\Pr\left[|D_{b,h} - E_{b,h}| > (1+\alpha)E_{b,h}\right] < e^{-\Omega(E_{b,h}\alpha^2)}$$

*Proof.* Let $X_i$ be the random variable which is 1 if the $(b-k+i)$th item has a bucket of height at least $i$, and 0 otherwise, and let $X = \sum_{i=0}^{k} X_i$. The result then follows immediately by applying Chernoff bounds to $X$. □

#### 3.3.4 Proof of Lemma 2

With the above proposition in place, we can now prove Lemma 2.

*Proof of Lemma 2.* Let $I_0 = [i - M\log^3 p + 1, i]$ and for $k \geq 1$ let

$$I_k = [\lceil i - (1+\alpha)\sigma(p)E_{(k-1)D}\rceil + 1, i],$$

and let $t_k$ denote the number of elements in the SkipList lying in $I_k$ with height at least $kD$. Define a Spray to be *viable at level $k$* if its $(\ell-k)$-prefix returns some element in $I_k$, and say that a Spray is *viable* if it is viable at every level. Intuitively, a Spray is viable at level $k$ if, after having finished walking at level $kD$, it ends up at a node in $I_k$. By Lemma 1, if a Spray is not viable at level $k$ for any $1 \leq k \leq \ell_p$, it will not return $x$ except with probability $p^{-\Omega(M\alpha^2\log^2 p)}$ over the choice of the SkipList, for all $k$. Thus, by a union bound, we conclude that if a Spray is not viable, it will not return $x$ except with probability $p^{-\Omega(M\alpha^2\log^2 p)}$ over the choice of the SkipList. It thus suffices to bound the probability that a Spray is viable.

Let $t_k$ be the number of elements in $I_k$ with height at least $kD$. The probability that the $kD$th level of any Spray lands in $I_k$ is at most $t_k/(M\log^3 p)$, since we choose how far to spray at level $kD$ uniformly at random. By Proposition 2 we know that except with probability $e^{-\Omega(\alpha^2(E_k+1))} = p^{-\Omega(M\alpha^2\log^2 p)}$, $I_k$ contains at most

$$\begin{aligned}&(1+\alpha)^2\sigma(p)E_{(k-1)D}2^{-kD} \\ &\quad = (1+\alpha)^2 M\sigma(p)\log^2 p\end{aligned}$$

elements with height at least $kD$. Hence,

$$\frac{t_k}{(M\log^3 p)} \leq (1+\alpha)^2\sigma(p)\frac{1}{\log p}$$

except with probability $p^{-\Omega(M\alpha^2\log^2 p)}$, for any fixed $k$. By a union bound over all $\log p/\log\log p$ levels, this holds for all levels except with probability $p^{-\Omega(M\alpha^2\log^2 p)}$. Thus, the probability that a Spray lands in $I_0$ after it completes but the traversal at the bottom of the list is

$$\left((1+\alpha)^2\sigma(p)\frac{1}{\log p}\right)^{\log p/\log\log p}$$

except with probability $p^{-\Omega(M\alpha^2\log^2 p)}$. If we choose $(1+\alpha)^2 = (1+\frac{1}{\log p})$ so that $\alpha = \sigma(p)^{1/2} - 1$, we obtain that since $(\log p)^{-\frac{\log p}{\log\log p}} = \frac{1}{p}$. Since $\sigma(p)^{\log p/\log\log p} \simeq 1$, and

$$\alpha^2\log^2 p = \left(\sqrt{\frac{\log p}{\log p - 1}} - 1\right)^2 \log^2 p \simeq \frac{1}{4},$$

it must be that with high probability, the fraction of Sprays that land in $I_0$ is asymptotically bounded by $p^{-1}$. Conditioned its $\ell$-prefix returning something in $I_0$, for the Spray to return $i$, it must further take the correct number of steps at the bottom level, which happens with at most a $\frac{1}{M\log^3 p}$ fraction of these Sprays. Moreover, if the $\ell$-prefix of the Spray does not return an element in $I_0$, then the Spray will not hit $i$, since it simply too far away. Thus $F_p(i,S) \lesssim \frac{1}{pM\log^3 p}$, as claimed. as claimed.

□

#### 3.3.5 Proof of Lemma 3

**Proof Strategy.** We wish to lower bound the probability of hitting the $i$th smallest item, for $i$ in some reasonable interval which will be precisely defined below. For simplicity of exposition in this section, we will assume that all the endpoints of the intervals we define here are integers are necessary. While this is not strictly true, the proof

is almost identical conceptually (just by taking floors and ceilings whenever appropriate) when the values are not integers and much more cumbersome.

Fix some index $i$. As in the proof of Lemma 2, we will again filter `Spray` by where they land at each level. By defining slightly smaller and non-overlapping regions than in the proof of Lemma 2, instead of obtaining upper bounds on the probabilities that a `Spray` lands at each level, we are instead able to lower bound the probability that a `Spray` successfully lands in the "good" region at each level, conditioned on the event that they landed in the "good" region in the previous levels.

Formally, let $I_0 = [i - \log^3 p, i - 1]$. Let $S$ be a spray, chosen randomly. Then if $i - \log^3 p \geq 0$, we know that if the $\ell$-prefix of $S$ returns an element in $I_0$, then $S$ has a $1/\log^3 p$ probability of stepping to $i$. Inductively, for all $k \leq \ell_p - 1$, we are given an interval $I_{k-1} = [a_{k-1}, b_{k-1}]$ so that $a_{k-1} \geq 0$. Notice that there are, except with probability $p^{-\Omega(M\alpha^2 \log^2 p)}$, at most $M \log^3 p$ elements in $[b_{k-1} - \frac{1}{1+\alpha} E_{kD}, b_{k-1}]$ with height $kD$, by Proposition 2.

Then, let $a_k = b_{k-1} - \frac{1}{1+\alpha} E_{(k-1)D}$ and $b_k = a_{k-1} - 1$, and let $I_k = [a_k, b_k]$. For all $0 \leq k \leq \ell_p - 1$, let $t_k$ be the number of elements in $I_k$ with height $(k+1)D$. Assume for now that $a_k \geq 0$. Then, if the $(k-1)$-prefix of $S$ returns an element $i$ in $I_k$, then every element of $I_{k-1}$ of height $kD$ by some walk of length at most $M \log^3 p$ at level $kD$, since there are at most $M \log^3 p$ elements of height $k \log\log p$ in the interval $[a_k, b_{k-1}]$ and $b_k < a_{k-1}$. Thus, of the `Spray`s whose $(k+1)$ prefixes return an element in $I_k$, a $t_k/(M \log^3 p)$ fraction will land in $I_{k-1}$ after walking at height $kD$. The following proposition provides a size bound on the $I_k$.

**Proposition 3.** *Let $s_k = b_k - a_k + 1$. For all $k \geq 2$, we have*

$$\left(\gamma_0 - \gamma_1 \frac{1}{\log p}\right) E_{kD} \leq s_k \leq \left(\gamma_0 + \gamma_1 \frac{1}{\log^2 p}\right) E_k$$

*with $\gamma_0 = \frac{\log p}{(\alpha+1)(\log p+1)}$ and $\gamma_1 = \frac{\alpha \log p + \alpha + 1}{(\alpha+1)(\log p+1)}$.*

*Proof.* Define $\xi_k$ to be the quantity so that $s_k = \xi_k E_k$. Clearly $\xi_0 = 1$, and inductively,

$$\begin{aligned} s_k &= \frac{1}{1+\alpha} E_k - s_{k-1} \\ &= \left(\frac{1}{1+\alpha} - \frac{\xi_{k-1}}{\log p}\right) E_k \end{aligned}$$

so

$$\xi_k = \frac{1}{1+\alpha} - \frac{1}{\log p} \xi_{k-1}.$$

Homogenizing gives us a second order homogenous recurrence relation

$$\xi_k = \left(1 - \frac{1}{\log p}\right) \xi_{k-1} + \frac{1}{\log p} \xi_{k-2}$$

with initial conditions $\xi_0 = 1$ and $\xi_1 = \frac{1}{1+\alpha} - \frac{1}{\log p}$. Solving gives us that

$$\xi_k = \gamma_0 + \gamma_1 \left(-\frac{1}{\log p}\right)^k.$$

Notice that $\xi_{2k+2} \leq \xi_{2k}$ and $\xi_{2k+3} \geq \xi_{2k+1}$ and moreover, $\xi_{2k+1} \leq \xi_{2k'}$ for any $k, k'$. Thus for $k \geq 2$ the maximum of $\xi_k$ occurs at $k = 2$, and the minimum occurs at $k = 1$. Substituting these quantities in gives us the desired results. □

Once this result is in place, we use it to obtain a lower bound on the hit probability.

**Lemma 4.** *There is some constant $A > 1$ which for $p$ sufficiently large can be chosen arbitrarily close to 1 so that for all $i \in [Ap \log^2 p, \frac{1}{1+1/\log p} Mp \log^3 p]$, we have $F_i(p) \gtrsim \frac{1}{Mp \log^3 p}$ with high probability.*

This statement is equivalent to the statement in Theorem 3.

*Proof.* The arguments made in Section A.3 are precise, as long as (1) every element of $I_{k-1}$ can be reached by a walk from anywhere in $I_k$ at level $k$ of length at most $M \log^3 p$, and (2) each $a_k \geq 0$. By Proposition 2, condition (2) holds except with probability $p^{-O(\alpha^2 M \log^2 p)}$. Moreover, each $a_k \geq 0$ is equivalent to the condition that $i \geq \log^3 p + \sum_{k=0}^{\ell_p - 1} s_k$, but by Proposition 3, we have that (except with probability $p^{-O(M\alpha^2 \log^2 p)}$) that

$$\sum_{k=0}^{\ell_p - 1} s_k \leq \left(\gamma_0 + \gamma_1 \frac{1}{\log^2 p}\right) \left(\sum_{k=0}^{\ell_p - 1} E_k\right).$$

For the choice of $\alpha = \frac{1}{\log p}$, the first term in this product can be made arbitrarily close to one for $p$ sufficiently large, and thus we have that except with probability $p^{-O(\alpha^2 M \log^2 p)}$,

$$\sum_{k=1}^{\ell_p - 1} s_k \quad \leq \quad AMp \log^2 p,$$

for some $A$ which can be made arbitrarily close to one for $p$ sufficiently large.

By Propositions 2 and 3, by a union bound, we have that except with probability $p^{-O(M\alpha^2 \log^2 p)}$,

$$t_k \geq 2^{-D} \left(\gamma_0 - \gamma_1 \frac{1}{\log p}\right) M \log^3 p,$$

for all $k$. Thus by the logic above, if we let $H_k$ denote the event that the $(k+1)$-prefix of the spray is in $I_k$, we have that the probability that the spray hits $i$ is

$$\begin{aligned} &\geq \Pr(\text{spray hits } i | H_0) \left(\prod_{k=1}^{\ell_p - 1} \Pr(H_{k-1} | I_k)\right) \Pr(H_{\ell_p - 1}) \\ &\geq \frac{1}{\log^3 p} \prod_{k=0}^{\ell_p - 1} \frac{t_k}{M \log^3 p} \\ &\geq \frac{1}{\log^3 p} \left(2^{-\lfloor \log\log p \rfloor} \left(\gamma_0 - \gamma_1 \frac{1}{\log p}\right)\right)^{\ell_p}. \end{aligned}$$

If we choose $\alpha = \frac{1}{\log n}$, then one can show that

$$\left(\gamma_0 - \gamma_1 \frac{1}{\log p}\right)^{\ell_p} \simeq 1,$$

so we conclude that $F_i(p) \gtrsim \frac{1}{p \log^3 p}$ with high probability. □

### 3.3.6 Proof of Corollary 1

We have shown so far that on a clean skip list, `Spray` operations act like uniformly random selection on a predictable interval $I$ near the front of the list of size tending to $Mp \log^3 p$. We justify here why this property is sufficient to guarantee that `Spray` operations execute in polylogarithmic time. A single `Spray` operation always takes polylogarithmic time, however, a thread may have to repeat the `Spray` operation many times. We show here that this happens with very small probability.

**Corollary 1.** *In the model described above,* `DeleteMin` *takes $O(\log^3 p)$ time in expectation. Moreover, for any $\epsilon > 0$,* `DeleteMin` *will run in time $O(\log^3 p \log \frac{1}{\epsilon})$ with probability at least $1 - \epsilon$.*

*Proof.* Recall a process has to retry if either (1) its `Spray` lands outside of $I$, or (2) the `Spray` collides with another `Spray` operation which has already removed that object. We know by Theorem 3 and more specifically the form of $I(p)$ given in Lemma 4 that (1) happens with probability bounded by $O(1/\log p)$ as $p \to \infty$ for each attempted `Spray` operation since Lemma 4 says that there are $O(p \log^2 p)$ elements before the start of $I(p)$, and Theorem 3 says that each is returned with probability at most $O(1/p \log^3 p)$, and (2) happens with probability upper bounded by the probability that we fall into set of size $p-1$ in $I$, which is bounded by $O(1/\log^3 p)$ for $p$ sufficiently large by Lemma 2. Thus, by a union bound, we know that the probability that `Spray` operations must restart is bounded by $O(1/\log p) \leq 1/2$ for $p$ sufficiently large. Each spray operation takes $\log^3 p$ time, and thus the expected time it takes for a `Spray` operation to complete is bounded by

$$\log^3 p \sum_{i=0}^{\infty} 2^{-1} = O(\log^3 p)$$

and thus we know that in expectation, the operation will run in polylogarithmic time, as claimed. Moreover, for any fixed $\epsilon > 0$, the probability that we will restart more than $O(\log(1/\epsilon)/\log\log p)$ times is at most $\epsilon$, and thus with probability at least $1-\epsilon$, we will run in time at most $O(\log^3 p \log(1/\epsilon)/\log\log p)$. □

## 4. Implementation Results

**Methodology.** Experiments were performed on a Fujitsu PRIMERGY RX600 S6 server with four Intel Xeon E7-4870 (Westmere EX) processors. Each processor has 10 2.40 GHz cores, each of which multiplexes two hardware threads, so in total our system supports 80 hardware threads. Each core has private write-back L1 and L2 caches; an inclusive L3 cache is shared by all cores.

We examine the performance of our algorithm on a suite of benchmarks, designed to test its various features. Where applicable, we compare several competing implementations, described below.

**Lotan and Shavit Priority Queue.** The SkipList based priority queue implementation of Lotan and Shavit on top of Keir Fraser's SkipList [14] which simply traverses the bottom level of the SkipList and removes the first node which is not already logically deleted. The logical deletion is performed using a `Fetch-and-Increment` operation on a 'deleted' bit. Physical deletion is performed immediately by the deleting thread. Note that this algorithm is *not* linearizable, but quiescently consistent. This implementation uses much of the same code as the `SprayList`, but does not provide state of the art optimizations.

**Lindén and Jonsson Priority Queue.** The priority queue implementation provided by Lindén et. al. is representative of state of the art of linearizable priority queues [21]. This algorithm has been shown to outperform other linearizable priority queue algorithms under benchmarks similar to our own. This algorithm is optimized to minimize compare-and-swap (CAS) operations performed by `DeleteMin`. Physical deletion is batched and performed by a deleting thread only when the number of logically deleted threads exceeds a threshold.

**Fraser Random Remove.** An implementation using Fraser's SkipList which, whenever `DeleteMin` would be called, instead deletes a random element by finding and deleting the successor of a random value. Physical deletion is performed immediately by the deleting thread. Although this algorithm has no ordering semantics whatsoever, we consider it to be the performance ideal in terms of throughput scalability as it incurs almost no contention from deletion operations.

**Wimmer et. al. $k$-Priority Queue.** The relaxed $k$-Priority Queue given by Wimmer et. al. [29]. This implementation provides a linearizable priority queue, except that it is relaxed in the sense that each thread might skip up to $k$ of the highest priority tasks; however, no task will be skipped by every thread. We test the hybrid version of their implementation as given in [29]. We note that this implementation does not offer scalability past 8 threads (nor does it claim to). Due to compatibility issues, we were unable to run this algorithm on the same framework as the others (i.e. Synchrobench). Instead, we show its performance on the original framework provided by the authors. Naturally, we cannot make direct comparisons in this manner, but the scalability trends are evident.

**SprayList.** The algorithm described in Section 2, which chooses an element to delete by performing a `Spray` with height $\lfloor \log p \rfloor + 1$, jump length uniformly distributed in $[1, \lfloor \log p \rfloor + 1]$ and padding length $p \log p/2$. Each thread becomes a *cleaner* (as described in Section 2.3) instead of `Spray` with probability $1/p$. Note that in these experiments, $p$ is known to threads. Through testing, we found these parameters to yield good results compared to other choices. Physical deletion is performed only by cleaner threads. Our implementation is built on Keir Fraser's SkipList algorithm [14], described in the Appendix, using the benchmarking framework of Synchrobench[9]. The code has been made publicly available [4].

### 4.1 Throughput

We measured throughput of each algorithm using a simple benchmark in which each thread alternates insertions and deletions, thereby preserving the size of the underlying data structure. We initialized each priority queue to contain 1 million elements, after which we ran the experiment for 1 second.

Figure 3 shows the data collected from this experiment. At low thread counts ($\leq$ 8), the priority queue of Lindén et. al. outperforms the other algorithms by up to 50% due to its optimizations. However, like Lotan and Shavit's priority queue, Lindén's priority queue fails to scale beyond 8 threads due to increased contention on the smallest element. In particular, the linearizable algorithms perform well when all threads are present on the same socket, but begin performing poorly as soon as a second socket is introduced above 10 threads. On the other hand, the low contention random remover performs poorly at low thread counts due to longer list traversals and poor cache performance, but it scales almost linearly up to 64 threads. Asymptotically, the `SprayList` algorithm performs worse than the random remover by a constant factor due to collisions, but still remains competitive.

To better understand these results, we measured the average number of failed synchronization primitives per `DeleteMin` operation for each algorithm. Each implementation logically deletes a node by applying a (`CAS`) operation to the deleted marker of a node (though the actual implementations use `Fetch-and-Increment` for performance reasons). Only the thread whose `CAS` successfully sets the deleted marker may finish deleting the node and return it as the minimum. Any other thread which attempts a `CAS` on that node will count as a failed synchronization primitive. Note that threads check if a node has already been logically deleted (i.e. the deleted marker is not 0) before attempting a `CAS`.

The number of `CAS` failures incurred by each algorithm gives insight into why the exact queues are not scalable. The linearizable queue of Lindén et. al. induces a large number of failed operations (up to 2.5 per `DeleteMin`) due to strict safety requirements. Similarly, the quiescently consistent priority queue of Lotan and Shavit sees numerous `CAS` failures, particularly at higher thread counts. We observe a dip in the number of `CAS` failures when additional sockets are introduced (i.e. above 10 threads) which we conjecture is due to the increased latency of communication, giving threads more time to successfully complete a `CAS` operation before a competing thread is able to read the old value. In contrast, the `SprayList` induces almost no `CAS` failures due to its collision avoiding design. The maximum average number of failed primi-

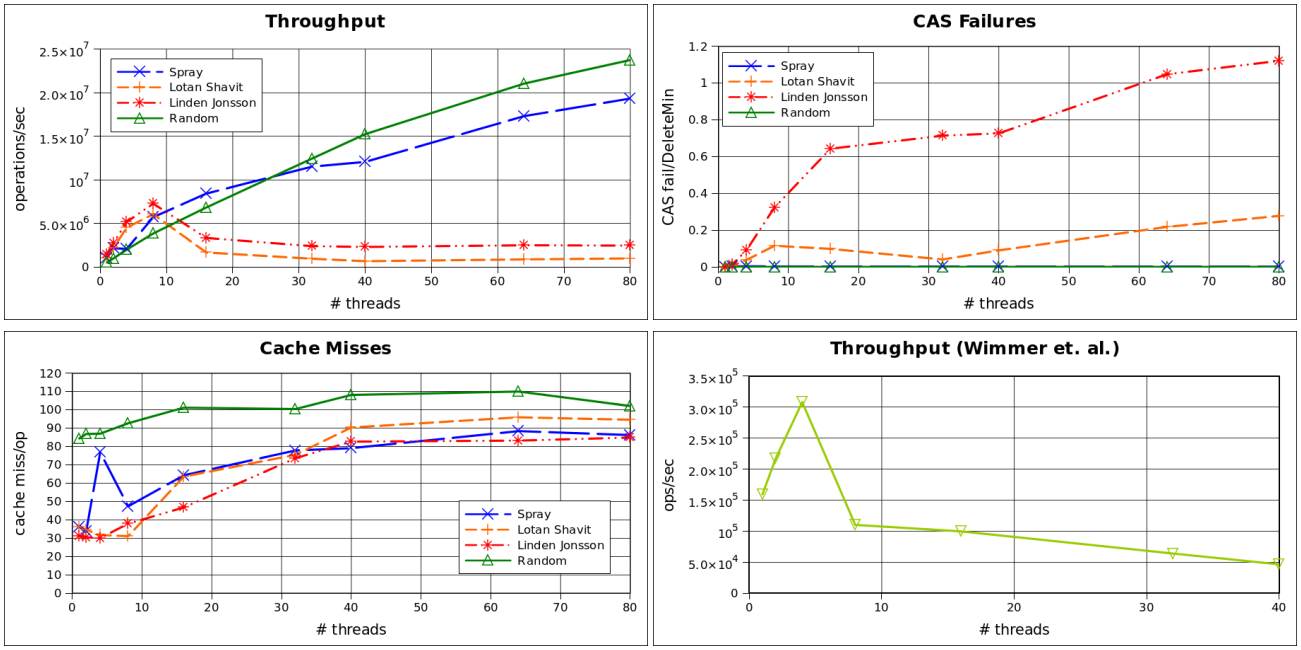


Figure 3: Priority Queue implementation performance on a 50% insert, 50% delete workload: throughput (operations completed), average CAS failures per DeleteMin, and average L1 cache misses per operation.

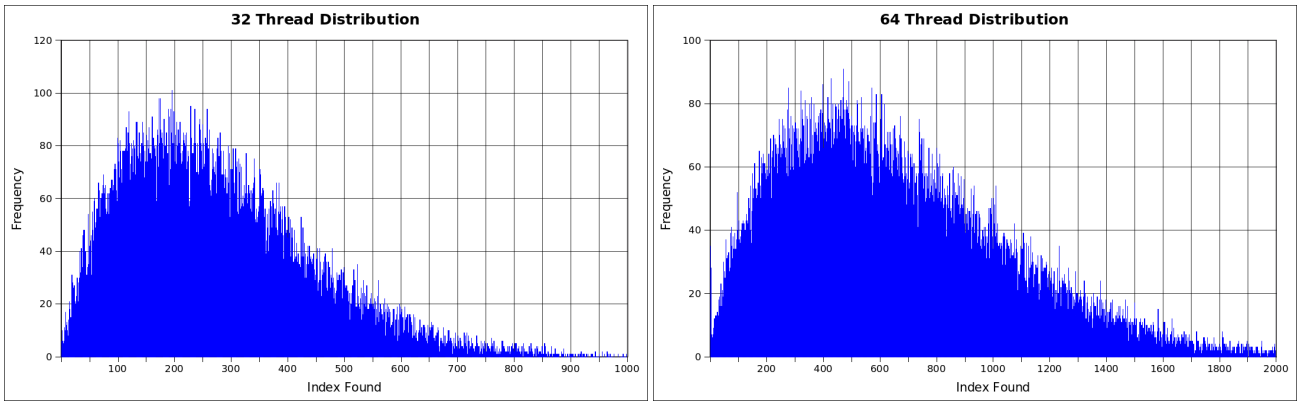


Figure 4: The frequency distribution of `Spray` operations when each thread performs a single `Spray` on a clean `SprayList` over 1000 trials. Note that the $x$-axis for the 64 thread distribution is twice as wide as for 32 threads.

tives incurred by the `SprayList` in our experiment was .0090 per `DeleteMin` which occurred with 4 threads. Naturally, the random remover experienced a negligible number of collisions due to its lack of ordering semantics.

Due to technical constraints, we were unable to produce a framework compatible with both the key-value-based implementations presented in Figure 3 and the task-based implementation of Wimmer et. al. However, we emulated our throughput benchmark within the framework of [29].

We implement tasks whose only functionality is to spawn a new task. Thus, each thread removes a task from the queue and processes that task by adding a new task to the queue. In this way, we measure the same pattern of alternating inserts and deletes in a task-based framework. As in the previous experiment, we initially populate the queue with 1 million tasks before measuring performance.

Figure 3 shows the total number of tasks processed by the $k$-priority queue of Wimmer et. al.[4] with $k = 1024$ over a 1 second duration. Similarly to the priority queue of Lindén et. al., the $k$-priority queue scales at low thread counts (again $\leq 8$), but quickly drops off due to contention caused by synchronization needed to maintain the $k$-linearizability guarantees. Other reasonable values of $k$ were also tested and showed identical results.

In sum, these results demonstrate that the relaxed semantics of `Spray` achieve throughput scalability, in particular when compared to techniques ensuring exact guarantees.

[4] We used the hybrid $k$-priority queue which was shown to have the best performance of the various implementations described [29].

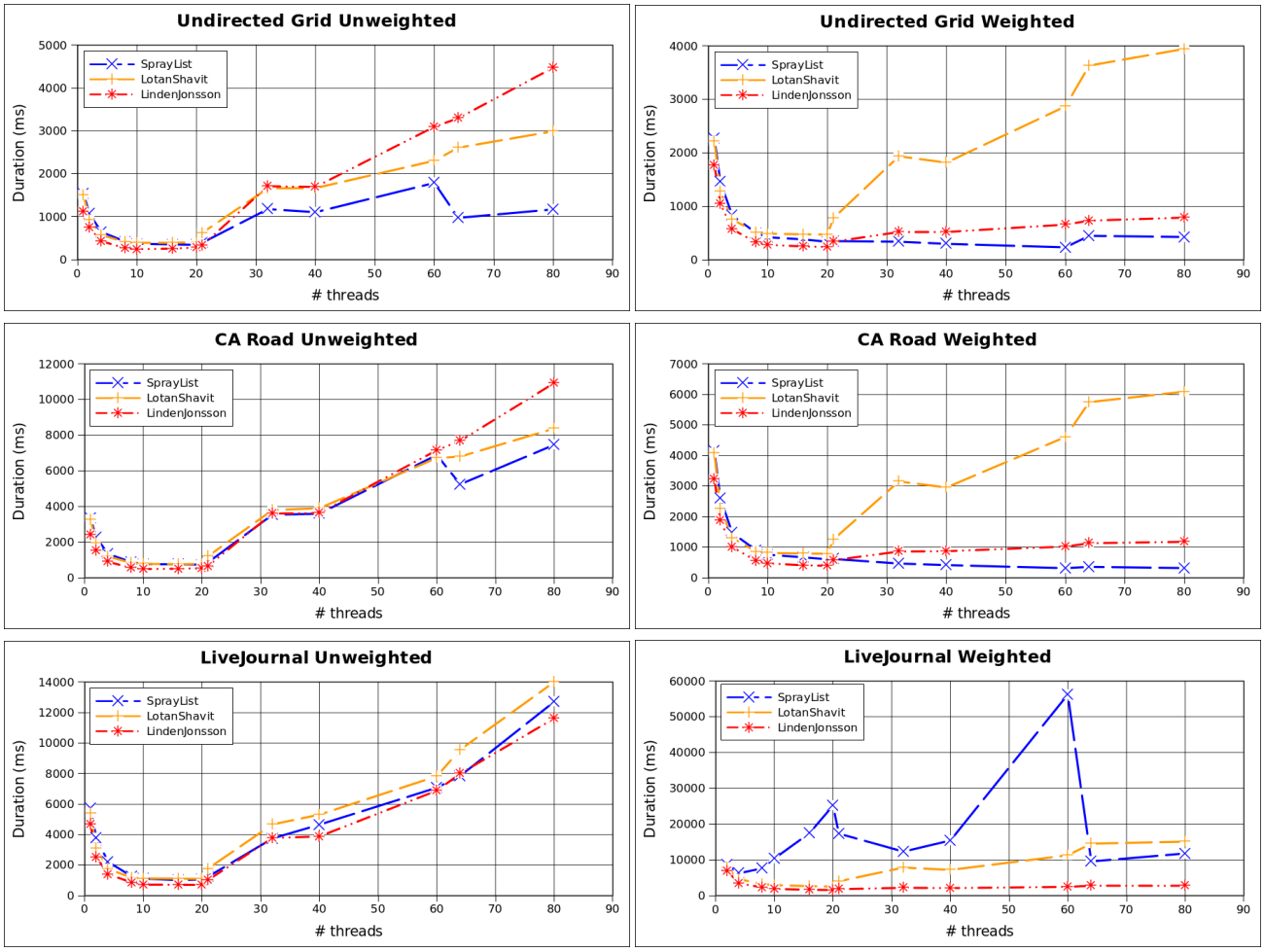


Figure 5: Runtimes for SSSP using each PriorityQueue implementation on each network (lower is better).

### 4.2 Spray Distribution

We ran a simple benchmark to demonstrate the distribution generated by the `Spray` algorithm. Each thread performs one `DeleteMin` and reports the position of the element it found. (For simplicity, we initialized the queue with keys $1, 2, \ldots$ so that the position of an element is equal to its key. Elements are not deleted from the `SprayList` so multiple threads may find the same element within a trial.) Figure 4 shows the distribution of elements found after 1000 trials of this experiment with 32 and 64 threads.

We make two key observations: 1) most `Spray` operations fall within the first roughly 400 elements when $p = 32$ and 1000 elements when $p = 64$ and 2) the modal frequency occurred roughly at index 200 for 32 threads and 500 for 64 threads. These statistics demonstrate our analytic claims, i.e., that `Spray` operations hit elements only near the front of the list. The width of the distribution is only slightly superlinear, with reasonable constants. Furthermore, with a modal frequency of under 100 over 1000 trials (64000 separate `Spray` operations), we find that the probability of hitting a specific element when $p = 64$ is empirically at most about .0015, leading to few collisions, as evidenced by the low `CAS` failure count. These distributions suggest that `Spray` operations balance the trade-off between width (fewer collisions) and narrowness (better ordering semantics).

### 4.3 Single-Source Shortest Paths.

One important application of concurrent priority queues is for use in Single Source Shortest Path (SSSP) algorithms. The SSSP problem is specified by a (possibly weighted) graph with a given "source" node. We are tasked with computing the shortest path from the source node to every other node, and outputting those distances. One well known algorithm for sequential SSSP is Dijkstra's algorithm, which uses a priority queue to repeatedly find the node which is closest to the source node out of all unprocessed nodes. A natural parallelization of Dijkstra's algorithm simply uses a parallel priority queue and updates nodes concurrently, though some extra care must be taken to ensure correctness.

Note that skiplist-based priority queues do not support the decrease-key operation which is needed to implement Dijkstra's algorithm, so instead duplicate nodes are added to the priority queue and stale nodes (identified by stale distance estimates) are ignored when dequeued.

We ran the single-source shortest path algorithm on three types of networks: an undirected grid ($1000 \times 1000$), the California road network, and a social media network (from LiveJournal) [1]. Since the data did not contain edge weights, we ran experiments with unit weights (resembling breadth-first search) and uniform random weights. Figure 5 shows the running time of the shortest

paths algorithms with different thread counts and priority queue implementations.

We see that for many of the test cases, the `SprayList` significantly outperforms competing implementations at high thread counts. There are of course networks for which the penalty for relaxation is too high to be offset by the increased concurrency (e.g. weighted social media) but this is to be expected. The LiveJournal Weighted graph shows a surprisingly high spike for 60 cores using the `SprayList` which is an artifact of the parameter discretization. In particular, because we use $\rfloor \log p \lfloor$ for the `Spray` height, the `Spray` height for 60 cores rounds down to 5. The performance of the `SprayList` improves significantly at 64 cores when the `Spray` height increases to 6, noting that nothing about the machine architecture suggests a significant change from 60 to 64 cores.

### 4.4 Discrete Event Simulation

Another use case for concurrent priority queues is in the context of Discrete Event Simulation (DES). In such applications, there are a set of events to be processed which are represented as tasks in a queue. Furthermore, there are dependencies between events, such that some events cannot be processed before their dependencies. Thus, the events are given priorities which impose a total order on the events which subsumes the partial order imposed by the dependency graph. As an example, consider $n$-body simulation, in which events represent motions of each object at each time step, each event depends on all events from the preceding time step. Here, a total order is given by the time step of each event, along with an arbitrary ordering on the objects.

We emulate such a DES system with the following methodology: we initially insert 1 million events (labelled by an ID) into the queue, and generate a list of dependencies. The number of dependencies for each event $i$, is geometrically distributed with mean $\delta$. Each event dependent on $i$ is chosen uniformly from a range with mean $i + K$ and radius $\sqrt{K}$. This benchmark is a more complex version of the DES-based benchmark of [21], which in turn is based on the URDME stochastic simulation framework [12].

Once this initialization is complete, we perform the following experiment for 500 milliseconds: Each thread deletes an item from the queue and checks its dependants. For each dependant, if it is not present in the queue, then some other thread must have already deleted it. This phenomenon models an inversion in the event queue in which an event is processed with incomplete information, and must be reprocessed. Thus, we add it back into the queue. We call this early deletion and reinsertion *wasted work*. This can be caused by the relaxed semantics, although we note that even linearizable queues may waste work if a process stalls between claiming an event and actually processing it.

This benchmark allows us to examine the trade-off between the relaxed semantics and the increased concurrency of `SprayLists`. Figure 6 reports the actual work performed by each of the competing algorithms, where actual work is calculated by simply measuring the reduction in the size of the list over the course of the experiment, as this value represents the number of nodes which were deleted without being reinserted later and can thus be considered fully processed. For each trial, we set $\delta = 2$ and tested $K = 100, 1000, 10000$.

As expected, the linearizable priority queue implementation does not scale for any value of $K$. As in the pure throughput experiment, this experiment also presents high levels of contention, so implementations without scaling throughput cannot hope to scale here despite wasting very little work.

On the other hand, the `SprayList` also fails to scale for small values of $K$. For $K = 100$, there is almost no scaling due to large amounts of wasted work generated by the loose semantics. However, as $K$ increases, we do start to see increased scalability, with $K = 1000$ scaling up to 16 threads and $K = 10000$ scaling up to 80 threads.

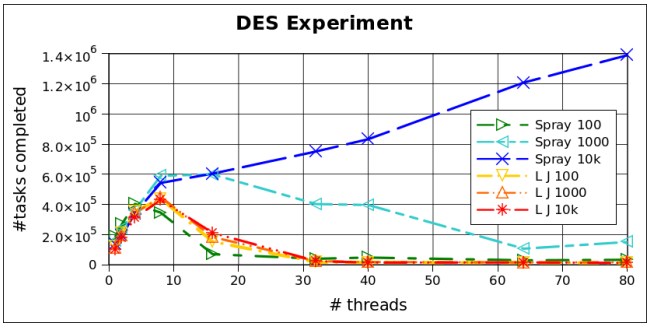


Figure 6: Work performed for varying dependencies (higher is better). The mean number of dependants is 2 and the mean distance between an item and its dependants varies between $100, 1000, 10000$.

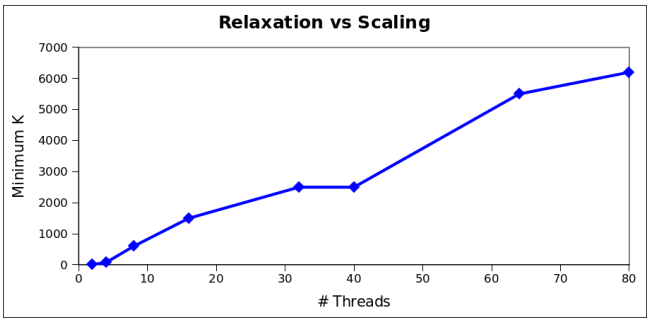


Figure 7: Minimum value of $K$ which maximizes the performance of the `SprayList` for each fixed number of threads.

To demonstrate the dependence of scalability on the distribution of dependencies, we measured the minimum value of $K$ needed to obtain maximum performance from a `SprayList` at each thread count. In particular, for each fixed value of $n$, we increased $K$ until performance plateaued and recorded the value of $K$ at which the plateau began.

Figure 7 reports the results of this experiment. We notice that the minimum $K$ required increases near linearly with the number of threads. Note that the "bumps" at 40 and 80 threads due to the dependence of `Spray` width only on $\lfloor \log_2 n \rfloor$ (so the minimum $K$ required will generally only increase at powers of 2). This plot suggests the required dependency sparsity in order for the `SprayList` to be a good choice of data structure for a particular application.

## 5. Discussion and Future Work

We presented a new design for a relaxed priority queue, which allows throughput scaling for large number of threads. The implementation weakens the strict ordering guarantees of the sequential specification, and instead provides probabilistic guarantees on running time and number of inversions. Our evaluation suggests that the main advantage of our scheme is the drastic reduction in contention, and that, in some workloads, the gain in scalability can fully compensate for the additional work due to inversions. We develop our technique on a lock-free SkipList, however a similar construct works for a lock-based implementation. Also, the relaxation parameters of our algorithm (spray height, step length) can be tuned depending on the workload.

An immediate direction for future work would be to tune the data structure for specific workloads, such as efficient traversals of

large-scale graphs. A second direction would be to adapt the spraying technique to obtain relaxed versions of other data structures, such as double-ended queues [17].

## References


[1] Stanford large network dataset collection. `http://snap.stanford.edu/data/index.html`. Accessed: Sept. 2014.

[2] Y. Afek, M. Hakimi, and A. Morrison. Fast and scalable rendezvousing. *Distributed Computing*, 26(4):243–269, 2013.

[3] D. Alistarh, J. Aspnes, S. Gilbert, and R. Guerraoui. The complexity of renaming. In *52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 718–727, Oct. 2011.

[4] D. Alistarh, J. Kopinsky, J. Li, and N. Shavit. Spraylist. `https://github.com/jkopinsky/SprayList`.

[5] D. Basin, R. Fan, I. Keidar, O. Kiselov, and D. Perelman. Cafe: Scalable task pools with adjustable fairness and contention. In *Proceedings of the 25th International Conference on Distributed Computing*, DISC'11, pages 475–488, Berlin, Heidelberg, 2011. Springer-Verlag.

[6] A. Braginsky and E. Petrank. A lock-free b+tree. In *24th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA '12, Pittsburgh, PA, USA*, pages 58–67, 2012.

[7] G. S. Brodal, J. L. Träff, and C. D. Zaroliagis. A parallel priority queue with constant time operations. *J. Parallel Distrib. Comput.*, 49(1):4–21, 1998.

[8] I. Calciu, H. Mendes, and M. Herlihy. The Adaptive Priority Queue with Elimination and Combining. *ArXiv e-prints*, Aug. 2014.

[9] T. Crain, V. Gramoli, and M. Raynal. A speculation-friendly binary search tree. *ACM SIGPLAN Notices*, 47(8):161–170, 2012.

[10] N. Deo and S. Prasad. Parallel heap: An optimal parallel priority queue. *The Journal of Supercomputing*, 6(1):87–98, Mar. 1992.

[11] D. Dice, Y. Lev, and M. Moir. Scalable statistics counters. In *25th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA '13, Montreal, QC, Canada*, pages 43–52, 2013.

[12] B. Drawert, S. Engblom, and A. Hellander. Urdme: a modular framework for stochastic simulation of reaction-transport processes in complex geometries. *BMC Systems Biology*, 6(76), 2012.

[13] F. Ellen, D. Hendler, and N. Shavit. On the inherent sequentiality of concurrent objects. *SIAM J. Comput.*, 41(3):519–536, 2012.

[14] K. Fraser. *Practical lock-freedom*. PhD thesis, PhD thesis, Cambridge University Computer Laboratory, 2003. Also available as Technical Report UCAM-CL-TR-579, 2004.

[15] T. L. Harris. A pragmatic implementation of non-blocking linked-lists. In *Proceedings of the 15th International Conference on Distributed Computing*, DISC '01, pages 300–314, London, UK, UK, 2001. Springer-Verlag.

[16] T. A. Henzinger, C. M. Kirsch, H. Payer, A. Sezgin, and A. Sokolova. Quantitative relaxation of concurrent data structures. In *Proceedings of the 40th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '13, pages 317–328, New York, NY, USA, 2013. ACM.

[17] M. Herlihy and N. Shavit. *The art of multiprocessor programming*. Morgan Kaufmann, 2008.

[18] M. Herlihy, N. Shavit, and M. Tzafrir. Hopscotch hashing. In *Proceedings of the 22nd International Symposium on Distributed Computing, DISC 2008, Arcachon, France*, pages 350–364, 2008.

[19] R. M. Karp and Y. Zhang. Parallel algorithms for backtrack search and branch-and-bound. *J. ACM*, 40(3):765–789, 1993.

[20] C. E. Leiserson, R. L. Rivest, C. Stein, and T. H. Cormen. *Introduction to algorithms*. The MIT press, 2001.

[21] J. Lindén and B. Jonsson. A skiplist-based concurrent priority queue with minimal memory contention. In *Principles of Distributed Systems*, pages 206–220. Springer, 2013.

[22] I. Lotan and N. Shavit. Skiplist-based concurrent priority queues. In *Parallel and Distributed Processing Symposium, 2000. IPDPS 2000. Proceedings. 14th International*, pages 263–268. IEEE, 2000.

[23] D. Nguyen, A. Lenharth, and K. Pingali. A lightweight infrastructure for graph analytics. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, SOSP '13, pages 456–471, New York, NY, USA, 2013. ACM.

[24] W. Pugh. Skip lists: a probabilistic alternative to balanced trees. *Communications of the ACM*, 33(6):668–676, 1990.

[25] W. Pugh. Concurrent maintenance of skip lists. 1998.

[26] P. Sanders. Randomized priority queues for fast parallel access. *Journal Parallel and Distributed Computing, Special Issue on Parallel and Distributed Data Structures*, 49:86–97, 1998.

[27] N. Shavit. Data structures in the multicore age. *Commun. ACM*, 54(3):76–84, 2011.

[28] H. Sundell and P. Tsigas. Fast and lock-free concurrent priority queues for multi-thread systems. *Journal of Parallel and Distributed Computing*, 65(5):609–627, 2005.

[29] M. Wimmer, D. Cederman, F. Versaci, J. L. Träff, and P. Tsigas. Data structures for task-based priority scheduling. In *Proceedings of the 19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 2014)*, 2014.