

Compact Parallel Hash Tables on the GPU

Citation for published version (APA):

Hegeman, S., Wöltgens, D., Wijs, A., & Laarman, A. (2024). Compact Parallel Hash Tables on the GPU. In J. Carretero, J. Garcia-Blas, S. Shende, I. Brandic, K. Olcoz, & M. Schreiber (Eds.), *Euro-Par 2024: Parallel Processing: 30th European Conference on Parallel and Distributed Processing, Madrid, Spain, August 26–30, 2024, Proceedings, Part II* (pp. 226-241). (Lecture Notes in Computer Science (LNC S); Vol. 14802). Springer. https://doi.org/10.1007/978-3-031-69766-1_16

Document license:

TAVERNE

DOI:

[10.1007/978-3-031-69766-1_16](https://doi.org/10.1007/978-3-031-69766-1_16)

Document status and date:

Published: 26/08/2024

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy




If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.



Compact Parallel Hash Tables on the GPU

Steef Hegeman¹(✉) , Daan Wöltgens^{2,3}, Anton Wijs² ,
and Alfons Laarman¹ 

¹ Leiden University,
Leiden, The Netherlands
{s.hegeman,a.w.laarman}@liacs.leidenuniv.nl

² Eindhoven University of Technology,
Eindhoven, The Netherlands
a.j.wijs@tue.nl

³ TNO, The Hague, The Netherlands
daan.woltgens@tno.nl



Abstract. On the GPU, hash table operation speed is determined in large part by cache line efficiency, and state-of-the-art hashing schemes thus divide tables into cache line-sized buckets. This raises the question whether performance can be further improved by increasing the number of entries that fit in such buckets. Known compact hashing techniques have not yet been adapted to the massively parallel setting, nor have they been evaluated on the GPU. We consider a compact version of bucketed cuckoo hashing, and a version of compact iceberg hashing suitable for the GPU. We discuss the tables from a theoretical perspective, and provide an open source implementation of both schemes in CUDA for comparative benchmarking. In terms of performance, the state-of-the-art cuckoo hashing benefits from compactness on lookups and insertions (most experiments show at least 10–20% increase in throughput), and the iceberg table benefits significantly, to the point of being comparable to compact cuckoo hashing—while supporting performant dynamic operation.

Keywords: Cuckoo hashing · Iceberg hashing · Quotienting · CUDA

1 Introduction

General purpose graphics processing units (GPUs) have been used to significantly speed up computations in many different domains. With thousands of processing cores, GPUs offer access to massive parallelism for everyone. However, the main GPU memory tends to be scarcer than the main memory available to CPUs. Moreover, memory access times are often the main performance bottleneck of GPU programs. Therefore, data structures that sparingly use GPU memory can positively affect both the amount of data that can be stored and the performance of a GPU program.

Memory-efficiency can be achieved by *quotienting*, a technique for reducing the storage required per key by using its storage location as information. This

was first used in practice in Cleary’s compact hash table [8]. This reduces the memory usage per table slot logarithmically in the number of slots. The parallelization of Cleary’s hashing scheme in [18] involves coarse locking. On the GPU, the (coarse) locking strategies of traditional CPU multicore algorithms do not perform well. The fastest GPU tables use atomic operations to directly insert keys into the table. We refer to this as *lockless*. In addition, it is fair to say that optimizing GPU hash table performance is about reducing the number of cache lines involved per operation [4].

Among the available GPU hash tables, adaptations of cuckoo hashing [1, 4, 15] and iceberg hashing [4, 7] have proven promising. In particular, the bucketed cuckoo table of [4], the slab hash table of [3] and DyCuckoo [13] are state-of-the-art. Indeed, [2, 7] confirm that at least in theory compact hashing with cuckoo or iceberg hashing is a good idea. A downside of these tables is however that they are more strict in terms of where a key can be stored. With cuckoo hashing, this can lead to situations where there are still empty slots in the table, but specific keys cannot be added to it. So here, one should not just consider the memory savings per slot, but also the expected *fill factor* of the table.

Moreover, not all hash tables offer the same guarantees. The bucketed cuckoo table of [4] for example is *static*, meaning that it only supports a mode of operation where the table is built once out of a batch of unique keys, and can then only be queried. As inserts temporarily move existing keys from the table to local memory, lookup operations performed during insertion may return false negatives, and concurrent insertion operations may cause keys to be stored in multiple slots. For the same reason, DyCuckoo [13], while supporting resizing, does not support concurrent insert operations. We say that a table supports *dynamic* operation if it supports concurrent combinations of lookups and writes. In many applications, dynamic operation is essential. For instance, in model checking [6, 19], fixpoint algorithms continuously check whether keys have been seen before and insert them if not.

We propose a compact, lockless GPU hash table that supports concurrent inserts. Our table is based on iceberg hashing, with a provably correct find-or-put operation. We provide an open source implementation in CUDA. We also combine the lockless GPU hash table based on cuckoo hashing from [4] with compact hashing for comparison. Synthetic benchmarks show that both tables benefit from compactness (10–20% speedup on lookups and insertions), while also halving or nearly quartering memory usage in many situations. Furthermore, the compact iceberg table performance is very close to the compact cuckoo one in static situations—a significant result, as the cuckoo table of [4] is, to the best of our knowledge, the fastest static GPU hash table to date, and the compact cuckoo table is comparable to it. This demonstrates the competitiveness of the compact iceberg table with the state-of-the-art.

In conclusion, we establish that when it comes to hash tables on the GPU, you can have your cake and eat it too: compact hashing can lead to both reduced memory usage as well as improved performance.

2 Background

2.1 Hash Tables

A hash table T is a data structure implementing a set or map from keys K to values V . Some hash tables are *stable*, meaning that the index at which a key is stored does not change. In this paper, we focus on hash tables representing sets, but the presented operations can be extended to a map implementation, by storing key-value or key-pointer pairs instead of keys, or, in the case of a stable table, by storing values in a separate array at the same index as the key.

Hash tables are typically optimized for FIND, PUT, and DELETE operations. We generally do not consider deletions (or treat them as if exceptional), as we are interested in the use of hash tables in search algorithms.

In this paper, our tables generally consist of an array of *buckets*. Each bucket contains one or more *slots*. Each slot is either unoccupied (has value EMPTY) or stores a key, possibly together with additional bookkeeping information. H hash functions $h_0, \dots, h_{H-1} : K \rightarrow T$ map keys to buckets in T . Given a hash function h_i , a key k can be stored in some slot with index j of the corresponding bucket $T[h_i(k)]$, i.e., $T[h_i(k)][j]$.

A table's *fill factor* is defined as the fraction of slots that are occupied. If key k is to be inserted into the table, but buckets indexed by $h_0(k), \dots, h_{H-1}(k)$ are full (all their slots are occupied), then there is typically no way to add k to the table and the table is considered full (an exception is the cuckoo hashing scheme below). A bigger table can then be allocated to store the set (*rehashing*) [15]. The larger H , the greater the expected fill factor a table achieves before rehashing is needed, but the longer lookups may take, as potentially all buckets $h_i(k), i < H$ need to be checked.

In which of the buckets $h_i(k)$ a key k is stored, and how the table is manipulated, is determined by the hashing *scheme*. The memory efficiency of a scheme depends on the expected fill factor that can be achieved, as well as the memory used per table slot. We discuss two schemes below.

2.2 Cuckoo Hashing

Insertion. Cuckoo hashing [9, 15, 17] achieves high fill factors in practice while limiting the slots in which a key may be placed, by moving keys between buckets during insertion. A yet to be inserted key may take the place of a key already in the table (*evicting* the original key), forcing it to be moved elsewhere.

Each slot either contains a dedicated value EMPTY, signaling it is unoccupied, or (in the case of cuckoo hashing) stores a pair (k, j) of key and hash function index $j < H$. When a slot containing (k, j) is evicted, the key k is in turn directed to bucket $h_{j+1 \bmod H}(k)$, evicting a slot there, until finally a key is directed to a bucket with an empty slot, in which case it takes the empty slot and the insertion is finished. If the chain of evictions exceeds a threshold length C , the insertion is aborted and the table is considered full. An optimization is to, instead of storing

a hash function index with the key, recover it from the key k found in bucket b as the first j with $h_j(k) = b$.

Lookup. A key k is only found in a bucket $h_{i+1}(k)$ if it was kicked out of bucket $h_i(k)$, so if deletions are forbidden, the $\text{FIND}(k)$ procedure only has to check indices $h_0(k), h_1(k), \dots$ in order up to the first bucket containing k or an unoccupied slot. (For rare deletion operations, additional ‘tombstone’ values can preserve this invariant, as discussed in, for example, [10].)

2.3 Iceberg Hashing

Iceberg hashing [7] divides the table into multiple *levels*. Each key is assigned to one large *primary bucket* in the first level by $h_0(k)$, and two smaller *secondary buckets* in the second level indicated by $h_1(k), h_2(k)$. There is an additional third level outside of the table structure, made up of linked lists. Iceberg hashing is stable: once a key is inserted into a slot, it is never moved to another [7].

Insertion. On insertions, a key k is placed in the primary bucket if it has unoccupied slots, otherwise it is placed in its secondary bucket with the most unoccupied slots. This choice aspect has significant impact on load balancing [5]. If both the primary and secondary buckets of k are full, then it gets sent to the third level, where k is inserted into the linked list indicated by $h_0(k)$. As level 3 can grow arbitrarily large, insertions can never fail. But to maintain performance, the table should be rehashed when the third level grows large.

Lookup. Again assuming deletions are prohibited, lookups can be performed as follows. First inspect the primary bucket. If it contains k , we are done, and if it does not contain k but some of its slots are unoccupied, we are done as well. Otherwise, both secondary buckets need to be inspected. If k is found in neither and one of the buckets has an unoccupied slot, then k is not in the table. If all three buckets are full and do not contain k , the linked list indicated by $h_0(k)$ needs to be searched for k .

Level Sizes. Not all buckets in the table have the same number of slots: the primary buckets are generally chosen to have more slots than the secondary ones [16]. Additionally, there are typically fewer secondary buckets than primary buckets [16]. So it might be instructive to think of an iceberg table as two hash tables and an array of linked lists.

The multi-level approach allows iceberg tables to achieve a high fill factor.¹ On the other hand, the more levels an operation needs to traverse, the longer it takes. If an operation only needs to work on the first two levels, it runs in constant time, but if level 3 also needs to be inspected, it does not. With proper tuning of the parameters, the expected spillage of level 1 to level 2 is small, and the spillage of level 2 to level 3 even smaller [7, 16]. In [16], they use large primary buckets (64 slots), and a 1 to 8 ratio in the size of the primary and secondary level. In practice, iceberg hashing achieves a high fill factor while maintaining near constant time performance (constant-time with high probability).

¹ In [7], definition of fill factor (space efficiency) is adapted to include level 3.

2.4 Compact Hashing

In [8], Cleary introduced a technique for compact hashing (now known as *quotienting* [7]), which reduces the storage space required per slot logarithmically in the number of buckets as follows.

Let T be a hash table with 2^N buckets, let $K = 2^M$ be the binary strings of length M , and let $\pi : K \rightarrow K$ be a permutation of keys. Assign key k to the bucket $a(k)$, where $a(k)$ is the number denoted by the first N bits of $\pi(k)$. This is also called the *address* of k . The remaining bits of $\pi(k)$, called the *remainder* $r(k)$, are then stored in a slot in this bucket. The key occupying a slot can thus be recovered by combining the address of the bucket that the slot is in and the remainder stored in the slot, followed by computing the inverse under π . In summary, $k = \pi^{-1}(a(k).r(k))$, where $.$ stands for concatenation of strings.

The remainders are only $M - N$ bits in length so per-slot and thus per-table memory savings are logarithmic in the total number of buckets.

Schemes using multiple hash functions h_i to assign keys to multiple buckets can be made compact by using multiple permutations π_i , giving rise to multiple address functions a_i and remainder functions r_i .

3 A Parallel Compact Iceberg Hash Table

We provide a lockless parallel compact iceberg algorithm. We focus on a parallel find-or-put operation which returns FOUND if the given key is in the table, and otherwise inserts the key (or returns FULL). With this operation, we can support ubiquitous fixpoint computations as the ones used in model checking [6, 19] and many other applications. Both cuckoo and iceberg hashing have proven promising for GPU applications, as discussed in the introduction. Because of the difficulty of realizing concurrent writes in cuckoo hashing, we opt for a compact iceberg table with a concurrent find-or-put operation.

Find-or-Put. Algorithm 1 gives our lockless parallel find-or-put procedure for compact iceberg hashing, called FOP. We use a table T_0 for the primary buckets, and T_1 for the secondary buckets, each compacted separately, so that $a_0(k)$ indicates a k 's primary bucket in T_0 , and $a_1(k), a_2(k)$ its secondary buckets in T_1 . The constants B_0 and B_1 are table parameters indicating the number of slots per primary bucket and per secondary bucket respectively.

Lines in the algorithm are not considered to be atomic, except for the compare-and-swap operation $\text{CAS}(a, b, c)$, which checks if the value of a is b , and if so, replaces it with c , all in one atomic operation. It returns TRUE if it did replace the value of a , and FALSE otherwise. (In particular, if $c \neq \text{EMPTY}$ and multiple threads simultaneously execute $\text{CAS}(a, \text{EMPTY}, c)$ on the same empty slot a , only one succeeds.) The reads in lines 2, 9, 11 are consequently atomic per-slot.

Correctness. A proof of correctness is given in the extended version [12]. The key idea is that when the algorithm attempts to insert into a slot, all earlier slots

Algorithm 1. Iceberg: lockless parallel find-or-put

```

FOP( $k$ )
1  while TRUE    ▷ Work on level 1
2     $b \leftarrow T_0[a_0(k)]$     ▷ Create a local snapshot
3    if ( $\exists i < B_0 : b[i] = r_0(k)$ ) return FOUND    ▷ Found  $k$ , we are done
4    if ( $\forall i < B_0 : b[i] \neq \text{EMPTY}$ ) break    ▷ Level 1 is full, go to level 2
5    else  $b$  has an empty slot—let  $b[i]$  be the first
6        if CAS( $T_0[a_0(k)][i], \text{EMPTY}, r_0(k)$ ) return PUT    ▷ Insertion attempt
7
8  while TRUE    ▷ Work on level 2
9     $b_1 \leftarrow T_1[a_1(k)]$ 
10   if ( $\exists i < B_1 : b_1[i] = (r_1(k), 0)$ ) return FOUND
11    $b_2 \leftarrow T_1[a_2(k)]$ 
12   if ( $\exists i < B_1 : b_2[i] = (r_2(k), 1)$ ) return FOUND
    ▷  $k$  is not in the table, try to insert it into the least full secondary bucket
13    $i \leftarrow 1$  if  $b_1$  is strictly less full than  $b_2$  else 2
14   if ( $\forall u < B_1 : b_i[u] \neq \text{EMPTY}$ ) return FULL
15   else  $b_i$  has an empty slot—let  $b_i[j]$  be the first
16       if CAS( $T_1[a_i(k)][j], \text{EMPTY}, (r_i(k), i)$ ) return PUT    ▷ Insertion attempt

```

are nonempty and have been read (see e.g. lines 5, 15). This allows the table to handle concurrent inserts with duplicate inputs.

Limitations. We limit ourselves to an iceberg table with level 1 and 2 tables—the difficulty in parallelizing iceberg hashing lies in the choice aspect of level 2. It should be noted that [4] does not truly separate the first and second level for its iceberg implementation, defying the basis of the theoretical analysis in [7]. We believe that the dynamic slab hash table of [3] could be adapted as a third level for the scheme below, forming a full 3-level iceberg table. We have found that already with the first two levels, good fill factors can be achieved in practice. We currently omit resizing and support for deletion operations, as a vast number of operations can already be supported with the given find-or-put operation. As discussed in for example [10], the table can be extended to support a delete operation using so-called ‘tombstone’ values, which preserve the invariant that non-empty slots occur consecutively. Using a stop-the-world approach, we believe resizing can also be implemented. In most applications on the GPU, it suffices however to simply claim all memory for the task at hand.

4 Implementation

4.1 Architectural Considerations

Architecture. We summarize the architecture of NVIDIA GPUs as described in the CUDA programming guide [14]. A GPU contains several multiprocessors, each capable of executing multiple *threads* (processes) in parallel. Threads are divided in groups of 32 called *warps*. Warps are then assigned to multiprocessors.

As a consequence of this warp-oriented architecture, if threads in a warp take different branches, the execution of these branches may be serialized. The highest performance is achieved if all threads in a warp execute the same lines of code. An upside of the tight coupling between threads in a warp is that they can efficiently communicate, and that their memory accesses can be *coalesced*: if threads in a warp access elements in the same cache line in parallel, the line is retrieved from memory only once, instead of once per thread.

In summary, to improve performance, threads in a warp should have as little branch divergence as possible, and should aim to access memory in the same cache line as often as possible.

Cooperative Work Sharing. For bucketed GPU hash tables, this is typically achieved by *warp-cooperative work sharing* [3,4]. Each thread receives an input key, but warps then cooperate, working together on one of their threads' keys at a time. When a bucket is inspected, each thread in the warp reads one of the bucket slots, together assessing the whole bucket in one (or few, depending on the bucket size) coalesced reads. CUDA allows for warps to be subdivided in smaller *cooperative groups*, which can be used when buckets have fewer than 32 slots. In summary, cooperative work sharing allows all threads to do useful work while decreasing the number of memory operations per thread.

Group Synchronization. In CUDA, each thread has a global *rank*, an index in the total number of threads. In a cooperative group, each thread also has a local *group rank*, from 0 to the group size (exclusive). Cooperative groups have several synchronization primitives, such as `shfl`, `any`, and `ballot`, which can be used to implement the following abstract procedures.

Let G be a group. $\text{SHUFFLE}_{G,s}(v)$ evaluates v in the thread with group rank s and returns the result. $\text{ANY}_G(P)$ is `TRUE` if and only if the predicate P evaluates to `TRUE` in any thread in the group, $\text{FIRST}_G(P)$ gives the group rank of the first thread in G in which P holds, or \perp if P is not true in any, and $\text{COUNT}_G(P)$ gives the number of threads in G in which P holds.

Global Synchronization. When one thread writes to memory (in our case, a bucket slot), there is no guarantee that this change is reflected in reads by other threads until explicit synchronization, unless this memory was written using atomic instructions. (Even then, volatile loads are required.) Of interest to us are `atomicCAS` and `atomicExch`, atomic `CAS` and `SWAP` operations, respectively.

4.2 Iceberg Find-or-Put

Algorithm 2 describes the cooperative find-or-put procedure. For simplicity, we assume that the primary bucket size B_0 must divide 32 (the size of a warp), and that the secondary buckets contain half that many rows.² Algorithm 3 implements the cooperative work sharing for inputs of a multiple of B_0 keys. The actual implementation supports input batches of any size.

² The actual implementation also supports smaller secondary buckets. Larger primary buckets could be implemented.

Implementation Notes. In the actual CUDA implementation, the reads use volatile loads, and there are some minor optimizations (filled slots are not read again, and it exploits that `atomicCAS` returns the read value of the target slot to avoid rereading the slot after a failed insertion attempt).

Algorithm 2. Iceberg find-or-put for cooperative group G with $|G| = B_0 = 2B_1$

```

COOPFOP( $k, G$ )
1   $rk \leftarrow$  my rank in  $G$ 
2  loop forever
3     $s \leftarrow T_0[a_0(k)][rk]$   $\triangleright$  Snapshot “my” bucket slot (coalesced read)
4    if  $\text{ANY}_G(s = r_0(k))$  return FOUND
5     $load \leftarrow \text{COUNT}(s \neq \text{EMPTY})$   $\triangleright$  Compute the number of filled slots
6    if  $load = B_0$  break  $\triangleright$  Level 1 is full, go to level 2
7    else  $\triangleright$  One of the threads tries to insert  $k$  into the first empty slot
8      if  $rk = load$ 
9         $s \leftarrow \text{CAS}(T_0[a_0(k)][rk], \text{EMPTY}, r_0(k))$ 
10     if  $\text{SHUFFLE}_{G,load}(s)$  return PUT  $\triangleright$  If it succeeds, we are done
11
12  subdivide  $G$  into  $G_1, G_2$  of even and odd threads respectively
13   $j \leftarrow (rk \bmod B_1) + 1$   $\triangleright$  Determine “my” subgroup
14   $rk' \leftarrow$  my rank in  $G_j$ 
15  loop forever
16     $s \leftarrow T_1[a_j(k)][rk']$   $\triangleright$  Subgroup  $j$  inspects the  $j$ th bucket (coalesced reads)
17    if  $\text{ANY}_G(r = (r_j(k), j - 1))$  return FOUND
18     $load \leftarrow \text{COUNT}_{G_j}(s \neq \text{EMPTY})$   $\triangleright$  Each subgroup calculates its bucket load
19     $load_1 \leftarrow \text{SHUFFLE}_{G,0}(load)$ 
20     $load_2 \leftarrow \text{SHUFFLE}_{G,1}(load)$ 
21     $\triangleright$  One thread tries to insert  $k$  into the least full bucket
22     $i \leftarrow 1$  if  $load_1 < load_2$  else 2
23    if  $load_i = B_1$  return FULL
24    else
25      if  $rk = i - 1$ 
26         $s \leftarrow \text{CAS}(T_1[a_i(k)][load_i], \text{EMPTY}, (r_i(k), i - 1))$ 
27        if  $\text{SHUFFLE}_{G,i-1}(s)$  return PUT

```

4.3 Permutations

We use simple permutation functions: one-round Feistel functions based on the hash family used in [4] for comparison purposes. Users of the CUDA implementation can easily supply their own permutations.

4.4 Parallel Compact Cuckoo Implementation

We give a compact cuckoo implementation, that is close to the bucketed cuckoo implementation of [4]. The main difference being the use of permutations instead of hash functions and the use of remainders, cf. Algorithm 4.

Algorithm 3. Find-or-put batch *keys* of size $n = mB_0$

```

FOP(keys,  $n$ )
1   $rk \leftarrow$  my global rank
2   $k \leftarrow \text{keys}[i]$ 
3  subdivide threads into groups of size  $B_0$ , let  $G$  be my group
4   $b \leftarrow \text{TRUE}$      $\triangleright$  I have work to do
5   $r \leftarrow \text{FULL}$ 
6  for  $0 \leq i < B_0$ 
7       $k' \leftarrow \text{SHUFFLE}_{G,i}(k)$ 
8       $r' \leftarrow \text{COOPFOP}(k', G)$ 
9      if  $rk = i$ 
10          $r \leftarrow r'$ 
11          $b \leftarrow \text{FALSE}$      $\triangleright$  I am done
12 return  $r$ 

```

A downside of cuckoo hashing is that it is not stable: keys move during insertions. This makes it difficult to define a performant find-or-put procedure: it could be that one process is checking whether k is in the table exactly when another process has decided to insert k' , and has in this process temporarily evicted k out of the table. We see only one way to avoid such situations, and that is to somehow synchronize all processes, so that no process works on the insertion-phase of a find-or-put while others are in the lookup-phase. However, on the GPU, synchronization of all parallel processes is very expensive.

Even if all processes are forced to finish their lookup-phase before one starts their insertion phase, there is yet another problem: multiple processes executing find-or-put for the same key k . After the lookup phase, they could all conclude that k has to be inserted into the table. During the insertion phase, after one process inserts k into its first bucket, an unrelated process might evict k from this bucket in order to insert a key k' , followed by one of the other processes inserting a new copy of k in the first bucket. This is another way in which multiple copies of a key could be inserted into the table. While there could still be solutions to this (locking, or using a prohibitive amount of metadata), we do not see a performant parallel find-or-put algorithm for cuckoo hashing. But as a static table, we can still make it compact.

Insertion. See Algorithm 4 for a parallel insertion algorithm for compact cuckoo. Note the inversion in line 11 to recover the evicted key. The CAS operation is atomic as in Algorithm 1. In addition the SWAP(a, b) operation atomically swaps the values of a and b . Again, the snapshot read in line 5 is atomic per-slot.

The CUDA implementation of Algorithm 4 uses the same warp-based work-sharing approach as the iceberg table.

Lookup. Look for $r_i(k)$ in buckets $a_i(k)$. As with non-compact cuckoo hashing, the insertion guarantees that if bucket $r_i(k)$ is not full, then buckets $r_j(k)$ with $j > i$ are completely empty (buckets are filled in order). Hence the search inspects

Algorithm 4. Compact cuckoo: parallel put

```

PUT( $k$ )
1   $j \leftarrow 0$ 
2  for  $c \in \{1, \dots, C\}$ 
3     $a \leftarrow a_j(k)$ 
4     $r \leftarrow r_j(k)$ 
5     $b \leftarrow T[a]$     ▷ Create a local snapshot
6    if  $\exists i < B: b[i] = \text{EMPTY}$     ▷ If there is an empty slot
7      if CAS( $T[a][i], \text{EMPTY}, (r, j)$ ) return PUT    ▷ Try to insert into it
8    else
9      choose  $i < B$     ▷ Pseudorandomly, or based on  $c, k$ 
10     SWAP( $(r, j), T[a][i]$ )    ▷ Atomically evict and claim a slot
11      $k \leftarrow \pi_j^{-1}(a.r)$     ▷ We now continue with the evicted key
12      $j \leftarrow j + 1 \bmod H$ 
13 return FULL

```

the buckets for k in order, and stops early if a bucket $a_i(k)$ is inspected that does not contain $r_i(k)$ and is not full (then k is not in the table).

4.5 CUDA Code

The full code, used in the benchmarks below, has been accepted as a conference artifact [11]. The latest version of our CUDA library is open source and available at <https://github.com/system-verification-lab/compact-parallel-hash-tables>. The cuckoo part is based partially on the second author’s master’s thesis [20]. For simplicity, we support only tables of which the number of slots (per level) is a power of 2, and we assume input keys are 64 bits wide. The code is set up so that these restrictions can be eliminated if so desired. Users can easily supply their own permutation functions.

The iceberg table uses `atomicCAS` (for CAS), and the cuckoo table also uses `atomicExch` (for SWAP). Both operations support word widths of 32, 64, and 128 bits, and `atomicCAS` additionally supports 16-bit words. Our cuckoo table thus supports slots of 32, 64 and 128 bits, and the iceberg table additionally supports slots of 16 bits.

Smaller slot sizes could be supported by using atomic instructions that are “too coarse”, likely at a performance cost. It is however important to be mindful of the global memory layout: the total memory usage of a bucket should divide the cache line size (128 bytes) lest some buckets end up being spread out over multiple cache lines, degrading performance.

5 Experimental Evaluation

5.1 Synthetic Benchmarks

We want to measure the performance impact of compact hashing. To this end, we set up tables (in various bucket combinations) with the same total number

of slots. Our original keys are wider than 32 bits, but we can store them in slots of 32 (cuckoo) or 16, 32 (iceberg) bits. We also benchmark the same tables with 64 bit slots that would fit the original keys. We can see the 64 bit tables as a baseline: they essentially behave as non-compact cuckoo and iceberg tables.³ Thus, our benchmarks show the impact of compact versus non-compact hashing in terms of runtime performance. Based on [4], we expect the tables with 32 slots per (primary) bucket in particular to benefit significantly, as the 64 bit versions use two cache lines per (primary) bucket and the compact tables one, or a half.

Input was taken from a set of uniformly drawn unique keys, taking duplicates from them as necessary. As in the benchmarks of [4], to keep the runtime manageable, we vary the permutations (associating keys to buckets) between measurements, instead of the input keys themselves.

We benchmark the largest table that can be stored on our GPU. Apart from the table, we also need to store input keys, and reserve memory for storing a.o. return values (FOUND, PUT, et cetera). With the 24GB memory limit of our RTX 4090, we end up with cuckoo tables of 2^{27} slots, and iceberg tables of 2^{27} primary and 2^{24} secondary slots.

Remark 1. Larger tables are possible with the use of CUDA unified memory [14, section 19.1.2], which allows for allocating more GPU memory than available, dynamically swapping memory to and from the host (CPU) RAM. Using unified memory, our RTX 4090 can work with tables of 2^{29} primary and 2^{26} secondary slots. In this case, the compact tables are $10\times$ faster than the non-compact ones. However, this is not a particularly fair comparison (the non-compact tables take at least twice as much memory, so they will require more swapping) we focus on the situation where the whole experiment fits in GPU memory.

If each primary bucket contains 32 of the 2^{27} primary slots, then there are 2^{22} primary buckets, and so compactness will shave 22 bits off of every key. If each primary slot is 16 bits wide, they can store remainders of at most 15 bits in length (one bit is required to indicate whether the slot is occupied or not), and so the primary level can store keys of at most $22 + 15 = 37$ bits. So we drew our input keys uniformly from $[0, 2^{37})$.

We benchmarked cuckoo tables of 8, 16, and 32 slots per bucket, with slot sizes of 32 (compact) and 64 (non-compact) bits. For iceberg, we benchmarked tables with 8, 16, and 32 slots per primary bucket—the secondary buckets having half the slots of the primary ones—for 16 bit primary slots with 32 bit secondary slots (compact),⁴ 32 bit primary slots with 32 bit secondary slots (compact), and 64 bit primary slots with 64 bit secondary slots (non-compact).

We reiterate that each cuckoo resp. iceberg table has the same number of slots, we only vary how they are divided into buckets and how much of the

³ Microbenchmarks have shown that the runtime of computing the permutations themselves is negligible, so this is a fair assumption.

⁴ We use 32 bit secondary slots with the 16 bit primary slots because the secondary slots have less compactness (as there are fewer secondary slots), and so the keys of width 37 would not fit in 16 bit secondary slots. They do fit in the 32 bit slots.

compactness is realized in practice (how much memory each slot consumes). For cuckoo, the compact 32 bit versions use half the memory of the non-compact 64 bit versions. For iceberg, the compact 32 bit versions use half the memory of the non-compact 64 bit versions, and the 16 bit versions use $\frac{9}{32}$ of the memory.

With other recent GPUs, the RTX 3090 and L40s, we have found results similar to the ones presented here. On the older RTX 2080 Ti, we have not.⁵

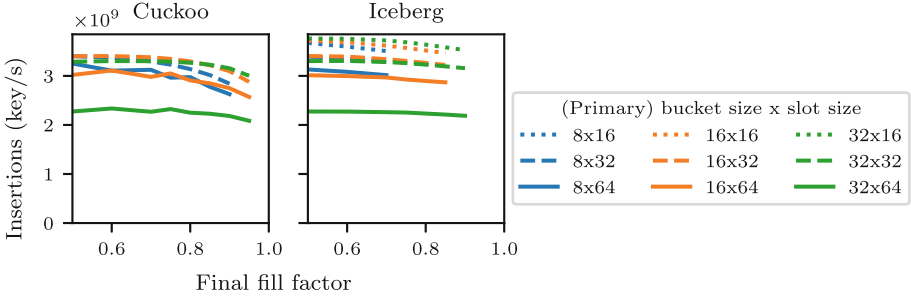


Fig. 1. PUT benchmark. The colors indicate bucket setup. The solid lines are non-compact tables (baseline), the **dashed lines** use half the memory (compact), and the **dotted lines** use roughly a quarter ($\frac{9}{32}$) the memory (extra compact).

5.2 Results

Insertion. Figure 1 shows the average throughput of filling the table to a certain fill factor with a batch of unique keys, measured for fill factors 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. (A static insertion benchmark.) The cuckoo tables with bucket size 16, 32 achieve fill factor 0.95. The highest fill factor achieved by the iceberg tables is 0.9, with bucket size 32. Both tables show similar performance, with a modest (5–10%) advantage for most compact 32 bit versions over 64 bit ones. The more compact 16 bit slots show an additional slight improvement, resulting in a 15–20% advantage over the 64 bit ones.

For the variants with 32 slots per (primary) bucket, there is a larger speedup. This is especially relevant for the iceberg table, as this only reaches fill factor 0.9 on this variant. The most compact (16 bit primary slots) version of this table is the fastest table in this benchmark, with a 60% performance increase over the non-compact version.

Lookup. Figure 2 shows FIND benchmark results, in which for each measurement the table is filled to a certain fill factor, and then queried for half as many unique keys as there are slots in the table. The tables with 32 slots per (primary) bucket benefit significantly from compactness, roughly doubling throughput over

⁵ On the RTX 2080 Ti, compactness shows only a negligible performance increase.

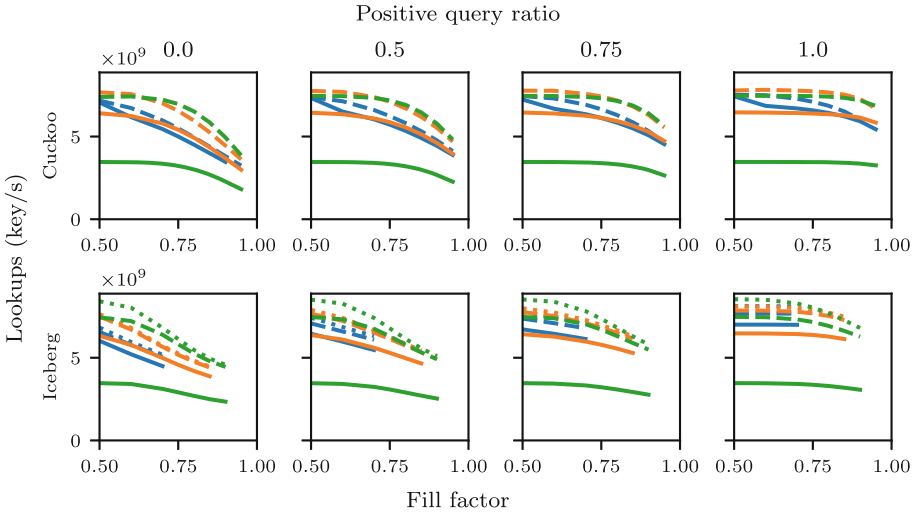


Fig. 2. FIND benchmark. Legend as in Fig. 1.

the non-compact version. For most other bucket sizes, compactness gives a rough 15–25% performance increase.

The best cuckoo table (compact, 32 slots per bucket) outperforms the best compact iceberg table (compact, 32 slots per bucket) by about 20–30% at higher load factors, especially when querying mostly keys that are not in the table. This can be explained by the fact that with iceberg tables, both secondary buckets are inspected if the primary bucket is full and does not contain k , while with cuckoo tables it can also happen that only two buckets are inspected (cf. Sect. 4.4).

Find-or-Put. We conducted a find-or-put benchmark, measuring throughput for various combinations of before and after fill factors. Before each measurement, the table was filled to the before fill factor. The FOP operation was then issued with as many input keys as there are slots in the table (2^{27} for cuckoo, $2^{27} + 2^{24}$ for iceberg), containing a mix (with duplicates) of keys not in the table and keys already in the table, such that after the operation, the table was filled exactly to the after fill factor.

To have a baseline for the iceberg table, we implemented an unsophisticated find-or-put operation for the cuckoo table that first sorts the input keys to detect duplicates (using the radix sort in Thrust, a library included with the CUDA toolkit), issues a FIND once per unique key, and then inserts one of each key not in the table with PUT.

Figure 3 shows the results. There is little difference between the cuckoo versions. The most compact 16 bit iceberg tables show a 10–20% speedup over the non-compact 64 bit variants, with a greater 60–100% speedup for the variant with 32 slots per primary bucket. The FIND-OR-PUT of the compact iceberg tables is more than 5 times faster than the unsophisticated cuckoo find-or-put.

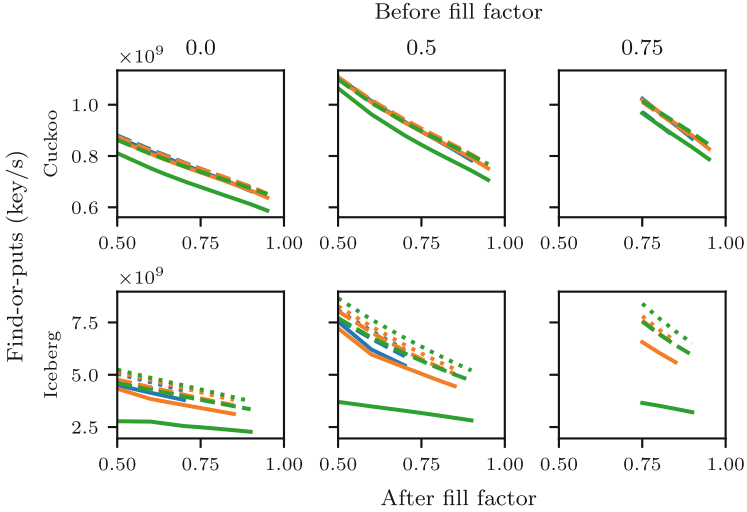


Fig. 3. FOP (find-or-put) benchmark. Legend as in Fig. 1. Note the scale difference.

5.3 Experiments with Real-World Data

Apart from synthetic data, we have also benchmarked the find-or-put operation against real-world data from a model checking application (HAVi from [19]). A single-threaded model checker was used to explore the model, and the sequence in which the nodes were visited forms our benchmark data.

The set contains about $2^{26.1}$ keys of width 24, about $2^{23.9}$ without duplicates. We measured the throughput of handling certain ratios of the data, for cuckoo tables of 2^{24} slots and iceberg tables of $2^{24} + 2^{21}$ slots.

See Fig. 4 for the results. The find-or-put of the fastest compact iceberg table is 8 times faster than the fastest baseline cuckoo find-or-put implementation on the RTX 4090. The most compact iceberg tables are competitive to their non-compact versions, with around 5–15% increase in throughput.

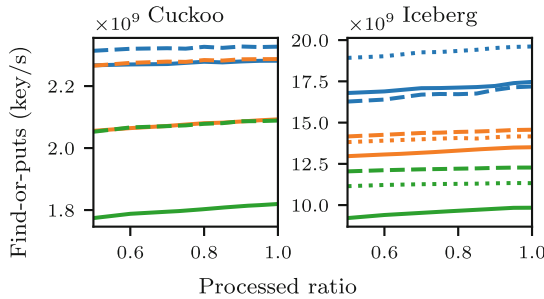


Fig. 4. FOP benchmark for the HAVi dataset. Legend as in Fig. 1. The throughput axes do not start at 0, so the differences in throughput are smaller than they appear.

6 Conclusion

On the GPU, compact hashing through quotienting not only saves precious GPU memory, it also modestly improves performance. (Compact) iceberg hashing provides a valid alternative to cuckoo hashing, with comparable FIND and PUT performance, while supporting an efficient, correct find-or-put—though our version with only two levels does not achieve as high load factors. A third level is worth considering, perhaps made of the slab lists of [16].

While we focused on modeling sets, similar techniques can be used to model key-value dictionaries. As this increases cache line strain (when storing the value with the key), compactness might be of even more importance here.

Artifact Availability. The code and data is available in the Zenodo repository [11].

References

1. Alcantara, D.A., Volkov, V., Sengupta, S., Mitzenmacher, M., Owens, J.D., Amenta, N.: Building an efficient hash table on the GPU. In: GPU Computing Gems Jade Edition, pp. 39–53. Elsevier (2012)
2. Arbitman, Y., Naor, M., Segev, G.: Backyard cuckoo hashing: constant worst-case operations with a succinct representation. In: IEEE FOCS (2010)
3. Ashkiani, S., Farach-Colton, M., Owens, J.D.: A dynamic hash table for the GPU. In: IEEE IPDPS, pp. 419–429 (2018)
4. Awad, M.A., Ashkiani, S., Porumbescu, S.D., Farach-Colton, M., Owens, J.D.: Analyzing and implementing GPU hash tables. In: APOCS. SIAM (2023)
5. Azar, Y., Broder, A.Z., Karlin, A.R., Upfal, E.: Balanced allocations (extended abstract). In: STC, STOC 1994, pp. 593–602. ACM (1994)
6. Baier, C., Katoen, J.: Principles of Model Checking. MIT Press, Cambridge (2008)
7. Bender, M.A., Conway, A., Farach-Colton, M., Kuszmaul, W., Tagliavini, G.: Iceberg Hashing: Optimizing Many Hash-Table Criteria at Once (2023). <https://doi.org/10.48550/arXiv.2109.04548>
8. Cleary, J.: Compact hash tables using bidirectional linear probing. IEEE Trans. Comput. **C-33**(9), 828–834 (1984)
9. Erlingsson, U., Manasse, M., McSherry, F.: A cool and practical alternative to traditional hash tables. In: DDS (2006)
10. Gunji, T., Eiichi, G.: Studies on hashing part-1: a comparison of hashing algorithms with key deletion. J. Inf. Process. **3**(1), 1–12 (1980)
11. Hegeman, S., Wöltgens, D., Wijs, A., Laarman, A.: Artifact of the paper: Compact Parallel Hash Tables on the GPU (2024). <https://doi.org/10.5281/zenodo.11638494>
12. Hegeman, S., Wöltgens, D., Wijs, A., Laarman, A.: Compact Parallel Hash Tables on the GPU (2024). <https://doi.org/10.48550/arXiv.2406.09255>
13. Li, Y., Zhu, Q., Lyu, Z., Huang, Z., Sun, J.: DyCuckoo: dynamic hash tables on GPUs. In: ICDE, pp. 744–755 (2021)
14. NVIDIA: CUDA C++ Programming Guide. Version 12.5 (2024). <https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html>
15. Pagh, R., Rodler, F.F.: Cuckoo hashing. J. Algorithms **51**(2), 122–144 (2004)

16. Pandey, P., et al.: IcebergHT: high performance hash tables through stability and low associativity. In: ACM on Management of Data, vol. 1 (2023)
17. Panigrahy, R.: Efficient hashing with lookups in two memory accesses. In: ACM SIAM Discrete Algorithms, SODA 2005, pp. 830–839. SIAM, USA (2005)
18. van der Vegt, S., Laarman, A.: A parallel compact hash table. In: Kotásek, Z., Bouda, J., Černá, I., Sekanina, L., Vojnar, T., Antoš, D. (eds.) MEMICS 2011. LNCS, vol. 7119, pp. 191–204. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-25929-6_18
19. Wijs, A., Osama, M.: A GPU tree database for many-core explicit state space exploration. In: Sankaranarayanan, S., Sharygina, N. (eds.) TACAS 2023. LNCS, vol. 13993, pp. 684–703. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-30823-9_35
20. Wöltgens, D.: Cleary-cuckoo: a compact parallelizable hash table. Master's thesis, Eindhoven University of Technology (2022)