GoFree: Reducing Garbage Collection via Compiler-Inserted Freeing

Haoran Peng*
University of Science and Technology of China
Hefei, China
hurrypeng@mail.ustc.edu.cn

Yu Zhang†
University of Science and Technology of China
Hefei, China
yuzhang@ustc.edu.cn

Michael D. Ernst
University of Washington
Seattle, USA
mernst@cs.washington.edu

Jinbao Chen
University of Science and Technology of China
Hefei, China
zkd18cjb@mail.ustc.edu.cn

Boyao Ding
University of Science and Technology of China
Hefei, China
via@mail.ustc.edu.cn

Abstract

In a memory-managed programming language, programmers allocate memory by creating new objects, but programmers never free memory. A garbage collector (GC) periodically reclaims memory used by unreachable objects. As an optimization based on escape analysis, some memory can be freed explicitly by instructions inserted by the compiler. This optimization reduces the cost of garbage collection, without changing the programming model.

We designed and implemented this explicit freeing optimization for the Go language. We devised a new escape analysis that is both powerful and fast ($O(N^2)$ time). Our escape analysis identifies short-lived heap objects that can be safely explicitly deallocated. We also implemented a freeing primitive that is safe for use in concurrent environments.

We evaluated our system, GoFree, on 6 open-source Go programs. GoFree did not observably slow down compilation. At run time, GoFree deallocated on average 14% of allocated heap memory. It reduced GC frequency by 7%, GC time by 13%, wall-clock time by 2%, and heap size by 4%. We made GoFree open-source.

CCS Concepts: • **Software and its engineering → Runtime environments**; **Automated static analysis**.

Keywords: escape analysis, explicit deallocation, Golang

ACM Reference Format:
Haoran Peng, Yu Zhang, Michael D. Ernst, Jinbao Chen, and Boyao Ding. 2025. GoFree: Reducing Garbage Collection via Compiler-Inserted Freeing. In *Proceedings of the 23rd ACM/IEEE International Symposium on Code Generation and Optimization (CGO '25), March 01–05, 2025, Las Vegas, NV, USA.* ACM, New York, NY, USA, 14 pages.
https://doi.org/10.1145/3696443.3708925

1 Introduction

Garbage collection (GC) [4, 12–14, 18, 24, 27, 29, 32, 33] frees developers from manual memory management, reducing the risk of memory-related bugs and simplifying coding. However, GC incurs the cost of extra memory usage and scanning time for dead objects [28, 35, 43].

*On a Ph.D. program at University of Washington at the time of publication.
†Corresponding author.

CGO '25, March 01–05, 2025, Las Vegas, NV, USA
2025. ACM ISBN 979-8-4007-1275-3/25/03
https://doi.org/10.1145/3696443.3708925


```
1  type BigType struct {
2    fat [10000000]int // Too
         big for stack
3    p   *int
4  }
5
6  func demo() *int {
7   var s []int
8   for i := 0; i < 10; i++ {
9     s = make([]int, 3)
10    // make() outlived by s
11  }
12
13  var b int
14  // b's address is stored
       into heap
15  var bigObj BigType // too big
16  bigObj.p = &b
17  var c int
18  // c's address is indirectly
       stored
19  var pc *int = &c
20
21  var d int // d escapes from
       function
22  var pd *int = &d
23  var ppd **int = &pd
24  *ppd = pc
25  // c's addr is stored into pd
26  var pd2 *int = *ppd
27  return pd2
28 }
```

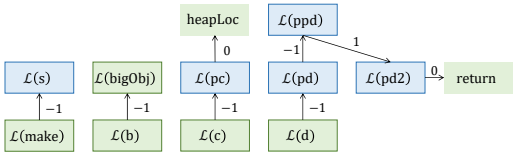

Figure 1: Go code and its escape graph. The escape graph determines whether a location (a node in the graph) is allocated on the stack (blue) or in the heap (green) by solving memory constraints. Virtual locations (nodes without frames: `heapLoc` and `return` are dummy locations for simplifying analysis that do not stand for specific memory units. Numbers on edges are explained in table 2.

Go [23] provides a lightweight escape analysis that supports the *stack allocation* optimization but not *explicit deallocation* (Section 2.1.2 explains these optimizations). Our system GoFree implements the explicit deallocation optimization.

Go's built-in escape analysis runs in $O(N^2)$ time, where N is the program's size (Throughout this paper, big O notation indicates the average case time complexity). To keep compilation fast, Go's escape analysis is *field-insensitive* and *flow-insensitive*. It conservatively assumes that anything whose address is assigned to an indirection escapes. For example, `*ppd = pc` on line 24 of fig. 1 causes an edge in Go's escape graph from `pc`'s node to a virtual "heap location" — which means "this address escapes to the heap" — rather than an edge connecting `ppd` and `pc`.

The escape analysis could be made more precise by providing it points-to information, which would enable creation of multiple normal edges rather than an edge to the virtual "heap location". Unfortunately, precise pointer analysis is generally $O(N^3)$ [2], which conflicts with Go's goal of fast compilation. Previous studies on Go escape analysis seldom

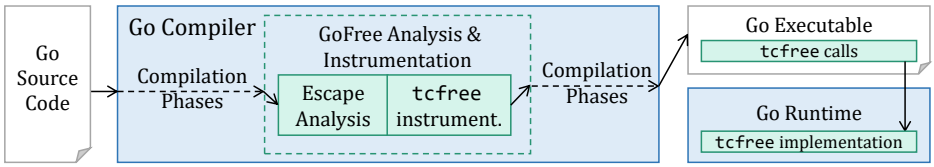

Figure 2: Architecture of GoFree. Blocks in blue are existing components of the Go compiler and runtime. Blocks in teal are new to GoFree. At compile time, GoFree inserts explicit deallocation (`tcfree` calls) in the executable. At run time, `tcfree` calls into the enhanced Go runtime. The GoFree enhancements are fewer than 2000 lines of code in total.

discuss its time complexity. For example, [40] finds an optimization opportunity in Go's handling of interface parameters. Other discussions [30] show code snippets where an intuitive improvement can be done, but without acknowledging that systematically addressing it would raise Go escape analysis's complexity to $O(N^3)$.

This paper presents GoFree, an $O(N^2)$ explicit deallocation analysis for Go that co-exists with Go's GC and its stack allocation optimization. Unlike approaches that increase precision at the cost of asymptotically greater run time, GoFree maintains $O(N^2)$ complexity by extracting relatively precise information from the original Go escape graph to support the explicit deallocation optimization. Unlike the stack allocation optimization, GoFree enables optimization across nested scopes and function calls.

GoFree consists of two main components (fig. 2): the *static analysis and instrumentation* component (section 4) and the *runtime* component for explicit deallocation (section 5).

The static analysis and instrumentation component performs completeness analysis (section 4.2), lifetime analysis (section 4.3), and enhanced inter-procedural analysis (section 4.4). The completeness analysis reveals which pointers' points-to sets computed from Go escape graph are complete, i.e. contains every variable that it can possibly point to at run time. The lifetime analysis determines which objects escape (outlive the pointer's lifetime); these cannot be explicitly deallocated. The instrumentation (section 4.5) inserts a `tcfree` primitive at the end of each complete, non-escaping, heap-allocated pointer's scope to deallocate the object.

A key part of the runtime support is our new `tcfree` primitive. For small objects allocated in thread-local caches, `tcfree` merely reverts the allocator pointer. For large objects allocated in the global heap, `tcfree` uses a 2-step asynchronous approach to avoid locking the entire heap (see fig. 9 and section 5). Unlike C's `free` primitive, `tcfree` does not guarantee successful collection. Whenever deallocation would be too costly (e.g., the object is in a non-thread-local place that requires locking to deallocate) or unsafe (e.g., GC is running, leading to a possible deallocation race), `tcfree` returns early leaving the object as if `tcfree` was never called. This strategy is safe: even if `tcfree` does nothing, GC will eventually deallocate the dead object.

GoFree differs from previous escape analysis work in two respects. First, GoFree sets compilation speed rather than precision as the priority to meet Go's design goal. We implemented a novel, precise deallocation analysis that outperforms Go's escape analysis without increasing its $O(N^2)$ time complexity. Second, GoFree handles Go features. One example is multiple return values: when the returned values have different allocation properties, a single per-method analysis and summary is inadequate. Another example is Go's *runtime-managed* dynamic data structures, such as slice and map, whose resizing is implemented in the runtime rather than as a library, and thus requires special support.

GoFree did not observably slow down compilation. At run time, GoFree deallocated on average 14% of allocated heap memory, and it reduced GC frequency by 7%, GC time overhead by 14%, wall-clock time by 2%, and heap size by 4%. The speedups apply to GoFree itself: when it is compiled by GoFree, it compiles other Go programs 1% faster.

In summary, our key contributions are as follows:

- An $O(N^2)$ static analysis for explicit deallocation.
- Runtime support for `tcfree`: explicit deallocation primitives that allow efficient deallocation.
- An open-source implementation of GoFree can be found at https://doi.org/10.6084/m9.figshare.26785357.

Our ideas apply to any runtime supported region-based or explicit deallocation. Some generalizable insights include:

- The static analysis can offload complexity to the runtime system. For example, `tcfree` accommodates double free in certain cases, so the static analysis can use `tcfree` even when it cannot prove no double free would occur.
- In a managed language, the `free` primitive is allowed to fail to free an object since GC will sweep it up in the end. This enables a low-cost best-effort `free` design.
- Identifying conservatism can indicate where results are precise (not conservative). Sound static analysis makes conservative assumptions; tracking their effect enables later analysis to improve the results or to utilize results that are already precise.

2 Related Work

2.1 Escape Analysis

Escape analysis [5, 15, 31, 37, 39] determines the scope and lifetime of objects. If fig. 3 is compiled without escape analysis, then both `make1` and `make2` will be allocated in the heap.

```
1 func analyses(n int) {
2   s1 := make([]int, 335) // make1
3   // ... use s1 ...
4   for i := 1; i < n; i++ {
5     s2 := make([]int, i) // make2
6     // ... use s2 ...
7   }
8 }
```

Figure 3: Go code snippet for comparing escape analyses

2.1.1 Compiler Optimizations Based on Escape Analysis.

Compiler optimizations built upon escape analysis can explicitly free an object when it is no longer accessed, rather than forcing GC to discover when it is no longer accessible. Or, it can allocate it physically onto the stack while it still logically resides in the heap.

Stack Allocation [10, 11, 19, 41]. Stack allocation identifies objects that are local to a method invocation and allocates them on the stack; popping the stack frees them, without any GC.

The Go compiler implements stack allocation. Each object of constant size that does not live beyond its scope (according to Go's built-in escape analysis) is allocated on the stack to avoid GC. In fig. 3, `make1` is allocated on the stack, while `make2` must be heap-allocated because its size is not a compile-time constant.

Region-based Memory Management [1, 6, 25, 38]. Region-based Memory Management allocates and frees memory in regions or arenas. No GC is required for any of the objects in the region because the entire region is returned to the free list when appropriate. However, no partial deallocation within the region is allowed before the whole region is freed.

Explicit Deallocation [9, 26, 36, 42]. Explicit deallocation identifies the last use of an object and automatically inserts an explicit free primitive after it, similar to what a C programmer would do. This is more fine-grained and flexible than stack or region-based allocation. It can free an object even if the object escapes from the original scope or has a variable size. However, it typically requires more precise information: points-to sets, flow sensitivity, and liveness.

Go does not support explicit deallocation, but GoFree does. In fig. 3, `make2` can be explicitly freed by inserting a `free` primitive call after line 6.

2.1.2 Precision and Complexity of Escape Analysis.

Fast Escape Analysis [19]. Fast escape analysis is an $O(N)$ algorithm supporting stack allocation. It only propagates escape properties among references and does not distinguish among `new`-ed objects. An object is stack-allocated iff the reference it is immediately bound to upon calling `new` does not escape. This simplification sometimes causes unnecessary heap allocation. Because the analysis does not provide any nontrivial points-to information, it is not capable of supporting explicit deallocation.

Connection Graph Based Escape Analysis [2, 10, 11, 26, 39]. A connection graph models the data flow among objects and their addresses, recording the effects of address fetches, assignments, indirect stores, and indirect loads. The algorithm propagates constraints across reference and object nodes. Such precision supports better stack allocation than Fast Escape Analysis and also supports explicit deallocation. However, as modeling indirect stores (e.g., `*p = q` in C, or `v.f = u` in Java) may generate up to $O(N)$ edges from a single statement, the analysis costs $O(N^3)$ (see section 3.2).

2.2 Related Memory Management Techniques

Deferred Reference Counting [16]. Deferred reference counting reduces reference counting by not reference-counting the stack. Go does not implement reference counting.

Scalar replacement [7, 37]. Scalar replacement decomposes a composite data structure into simpler ones, making them more likely to be eligible for stack allocation. This is helpful in Java, where an object and its components can only be accessed via references. Go uses explicit pointers and thus allows access to an object by pointer or value, so scalar replacement is impossible. GoFree implements a similar optimization for multiple return values (section 4.6.3).

3 Background about Go

3.1 Go memory management

Go is a memory-managed language. All objects are conceptually allocated in the heap [22] and freed by GC. Physically, they can be allocated in the heap or on the stack, but they must respect the following two memory invariants [21]:

- **Invariant 1**: A pointer to a stack object is not stored in the heap.
- **Invariant 2**: A pointer to a stack object does not outlive that object.

3.2 Go Escape Analysis

Go's escape analysis balances compilation speed and execution performance. It simplifies the connection graph approach by removing flow-sensitivity, field-sensitivity, and the tracking of indirect stores (e.g., `*p = q`). Conservative handling of these cases results in some precision loss compared to the connection graph approach, but it still performs better than Fast Escape Analysis. This simplified connection graph is called *escape graph*.

With indirect stores omitted, each Go statement can generate at most a constant number of nodes and edges for the graph, so the number of nodes and edges in the graph are both $O(N)$. Go then walks the graph and propagates properties using a modified Bellman-Ford algorithm [3] which has $O(N^2)$ time complexity on the sparse graph.

Section 4.1 formalizes Go's escape analysis.

Table 1: Escape properties used in GoFree

Property[Def#]	Cost[‡]	Description
LoopDepth[4.3†]	$O(N)$	layers of loops that the location is in
HeapAlloc[4.10†]	$O(N^2)$	location must be heap-allocated
Exposes[4.11]	$O(N^2)$	loc. may cause untracked modification to its referent
Incomplete[4.12]	$O(N^2)$	loc.'s value may be changed by untracked dataflow
DeclDepth[4.13]	$O(N)$	layers of scopes that the location is in
OutermostRef[4.14]	$O(N^2)$	the smallest scope that covers location's lifetime
Outlived[4.15]	$O(N^2)$	loc. has shorter lifetime than any object it points to
PointsToHeap[4.16]	$O(N^2)$	location may point to at least one heap object
ToFree[4.17]	$O(N^2)$	location is qualified to be deallocated by `tcfree`

† These two properties come from Go's original escape analysis.
‡ $O(N)$ arises from an AST scan. $O(N^2)$ arises from propagation.

3.3 Go's Heap Allocator and Garbage Collector

Go uses a thread-caching, size-segregated, free-list allocator called TCMalloc [34] to support high-concurrency environments, eliminating global locking for most allocations. After requesting memory pages from the OS, TCMalloc divides them into arena chunks of different small object sizes called *mspans*. Each thread may cache some mspans in its exclusively owned *mcache* so that most allocations from its mspans are lock-free unless the *mspan* runs out and needs to request another *mspan* from the OS.

Go's garbage collector accommodates TCMalloc and concurrent environments. It adopts a non-moving strategy because Go allows the explicit use of pointers by programmers.

4 Static Analysis and Instrumentation

This section details GoFree's explicit free analysis and instrumentation. Section 5 discusses runtime support.

Section 4.1 introduces the original $O(N^2)$ Go escape analysis framework. The two key analyses of GoFree are completeness analysis (section 4.2) and lifetime analysis (section 4.3); their extra constraints fit in the $O(N^2)$ framework. Section 4.4 extends GoFree to support inter-procedural analysis (IPA) that discovers more optimization opportunities across function calls. Section 4.5 discusses insertion of `tcfree` calls. Section 4.6 discusses how GoFree supports and utilizes Go's modern language features.

4.1 Formalization of the Go Escape Analysis

This section formally describes the original Go escape analysis, an $O(N^2)$ algorithm built for stack allocation optimization. It has two steps: building an escape graph that models program dataflow, and solving escape properties (see table 1 for a complete list) based on constraints on the graph.

Go's escape analysis yields a memory allocation scheme that is correct (each variable in Go exists as long as there are references to it) and strives for efficiency (put as many objects on the stack as possible). It first builds a directed weighted graph (called the escape graph) to model the data flow among objects. AST nodes that allocate memory are represented by escape graph nodes called locations, and dataflow relations are represented by escape graph edges.

Table 2: Go escape graph edges. $Derefs(e)$ is the weight of edge e.

code	edge	meaning of edge
`p = *q`	`q` $\xrightarrow{1}$ `p`	`p` may equal `*q`
`p = q`	`q` $\xrightarrow{0}$ `p`	`p` may equal `q`
`p = &q`	`q` $\xrightarrow{-1}$ `p`	`p` may equal `&q`
`*p = q`	`q` $\xrightarrow{0}$ `heapLoc`	`q`'s value may be in the heap

Definition 4.1 (Escape Graph)**.** An escape graph is a directed weighted graph, $EG = (L, E)$, where

- L is a set of locations defined in definition 4.2. We use l and m to represent locations in this paper.
- E is a set of edges defined in definition 4.4. We use e and $\langle l, m\rangle$ to represent edges in this paper.

Definition 4.2 (Location)**.** A location $l \in L$ is a vertex in the escape graph, which represents some storage space. When we say $l = \mathcal{L}(n)$, we mean that l represents the storage space created (explicitly or implicitly) by a (declaration or expression) node n in the AST. That is, $\mathcal{L}$ maps an AST node to its corresponding location.

Dummy locations are introduced to simplify the escape graph. Location `heapLoc` is a global constant that represents a heap location. A per-function location `return` represents the memory space of the function's return value.

Definition 4.3 (LoopDepth)**.** $LoopDepth(l) \in \mathbb{Z}$ records the loop depth at the declaration of variable n, where $l = \mathcal{L}(n)$. For dummy locations, $LoopDepth(\texttt{heapLoc}) = -1$, and $LoopDepth(\texttt{return}) = -1$.

Definition 4.4 (Edge)**.** $e = \langle l_1, l_2\rangle \in E$ is a directed weighted edge in the escape graph.

Definition 4.5 (Derefs)**.** $Derefs(e) \in \mathbb{Z}$ is the weight of edge e, representing the dereference count from l_1 to l_2.

Table 2 shows how edges model assignments. Indirect store `*p = q` is not further tracked, to avoid generating $O(N)$ edges for this single statement and degrading Go's escape analysis to $O(N^3)$. This simplification is safe because it conservatively indicates that q's value could go into the heap in the worst case.

Go's escape analysis is *field-insensitive*. When abstracting an object or an array as a location, all fields of the object or all elements of the array are represented by one single location.

Go's escape analysis is *flow-insensitive*. The order and scope of statements do not affect the escape graph. Since each statement generates a constant number of locations and edges, $|L| = O(N)$ and $|E| = O(N)$.

Using the rules introduced above, fig. 1 shows a code example and its escape graph. Once the escape graph is complete, Go creates and solves constraints on escape properties.

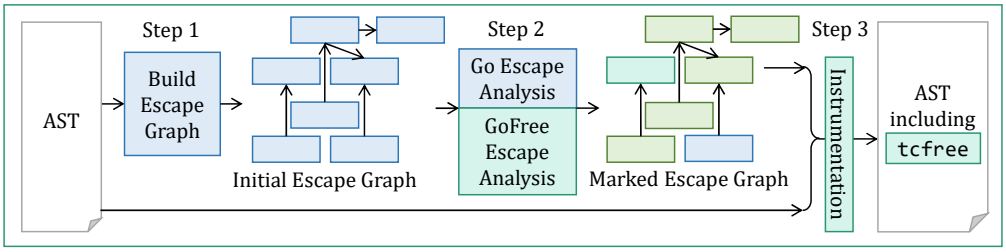

Figure 4: GoFree static analysis and instrumentation. GoFree reuses Go's original escape graph. After escape analysis, nodes in the graph are marked with different properties (such as *HeapAlloc*) that satisfy both Go's and GoFree's constraints. `tcfree` calls are then inserted into the AST according to these properties. Blue and green boxes are in the original Go system. Teal boxes are new to GoFree.

Definition 4.6 (Holds)**.** $m \in Holds(l)$ means that l can possibly contain the value of m, the address of m, and/or the value of a location that m points to (if m is a pointer). In the context of $Holds(l)$,

- $l \in L$ is called the root location,
- each $m \in Holds(l)$ is called a leaf location.

$|Holds(l)| \leq |L| = O(N)$, so by walking the escape graph in the reverse direction of edges, Go computes $Holds(l)$ in $O(N)$ time. $Holds(l)$ may be *incomplete*, missing some locations that also hold l because the escape graph does not track indirect stores.

Definition 4.7 (TrackDerefs)**.** $TrackDerefs(l_0 l_1 l_2 ... l_n) \in \mathbb{Z}$ is how many times value l_n is dereferenced when it is obtained via track $l_0 l_1 l_2 ... l_n$. Track is loop-free. $TrackDerefs(l_0 ... l_n)$ is computed by adding each $Derefs(l_i)$ value when walking the track in reverse, maintaining a lower bound 0 before each addition.

- $TrackDerefs(l_{n-1} l_n) = Derefs(\langle l_{n-1}, l_n \rangle)$
- $TrackDerefs(l_i l_{i+1} ... l_n) = max(0, TrackDerefs(l_{i+1} ... l_n)) + Derefs(\langle l_i, l_{i+1} \rangle)$

$TrackDerefs(l_0 ... l_n) = -1$ infers that l_0's address can possibly be obtained by l_n via this track. $TrackDerefs(l_0 ... l_n) \geq 0$ infers that l_0's value, or some value retrieved by dereferencing it one or more times can be obtained by l_n via this track. $TrackDerefs(l_0 ... l_n) \geq -1$ always holds.

Definition 4.8 (MinDerefs)**.** $MinDerefs(m, l)$ is the minimum value of $TrackDerefs(t)$ via any track t from m to l. It is defined iff $m \in Holds(l)$.

$$MinDerefs(m, l) = \min_{t=m..l} TrackDerefs(t)$$

Go computes the smallest dereference count because a variable can be a plain object (without pointer fields), a pointer (or an object containing pointer fields), or even a pointer to a pointer (we call this an order-2 pointer). Two different tracks with the same source and destination may have different $TrackDerefs()$. For example, `bigObj` in fig. 1 acts as a order-0 pointer (i.e., a value) with field `fat` (line 2) and a order-1 pointer with field `p` (line 3). Taking the minimum value of dereferences considers the objects as its highest-order role, and thus models in the escape graph its ability to transport any lower-order pointers.

Definition 4.9 (PointsTo)**.** $PointsTo(l) \subseteq L$ represents the points-to set of location l, where

$$m \in PointsTo(l) \text{ iff } MinDerefs(m, l) = -1.$$

$m \in PointsTo(l)$ means that l's value may be the address of m, similar to the result of points-to analysis. A queue-optimized Bellman-Ford algorithm [3, 8, 17] runs in $O(N)$ average time on such a sparse graph with limited edge weight. $PointsTo(l)$ can be incomplete due to the simplification of indirect stores. Section 4.2 on completeness analysis further discusses this.

Definition 4.10 (HeapAlloc)**.** $HeapAlloc(l)$ is true if location l must be heap-allocated. $HeapAlloc(l)$ is true if

- l = `heapLoc`, or
- l = `return`, or
- $\exists m \,.\, l \in PointsTo(m) \wedge HeapAlloc(m)$, or
- $\exists m \,.\, l \in PointsTo(m) \wedge LoopDepth(m) < LoopDepth(l)$.

where l is in the same function as m.

Go's escape analysis (fig. 5) finds a minimum solution for constraints. Properties propagate only from root locations to leaf locations, similar to the slack operation in the Bellman-Ford algorithm. Inspired by this, Go's escape analysis keeps

```
1  func walkall(EG = (L, E)):
2    work := UniqueQueue(copy(L))
3    while len(work) > 0: // O(N) repetitions
4      root := work.pop()
5      for leaf in Holds(root): // O(N) repetitions
6        leafUpdated := applyConstraints(root, leaf)
7        if leafUpdated:
8          work.push(leaf)
9        // Extension from GoFree.
10       // rootUpdated := applyConstraints(leaf, root)
11       // if rootUpdated:
12       //   work.push(root)
13       //   break
```

Figure 5: Go's $O(N^2)$ property propagation algorithm. `applyConstraints()` updates the properties (see table 1) of its second argument according to formulas shown in their definitions. For formulas that involve relationships between two nodes, the first argument provides more information of the corresponding root (or leaf) in the root-leaf pair.

Table 3: Points-to sets in different escape analyses. GoFree uses the Go escape graph, but analyzes which points-to sets are complete. In this example, GoFree recognizes that *PointsTo*($\mathcal{L}$(pd2)) is incomplete and refuses to explicitly deallocate pd2.

Method	Fast Esc. Analysis	Go esc. graph	Conn. graph
Time complexity	$O(N)$	$O(N^2)$	$O(N^3)$
Omitted dataflow	`*ppd = pc;` `pd2 = *ppd`	`*ppd = pc`	none
Conservatively modeled as	`{heap} = pc;` `{heap} = pd2`	`{heap} = pc`	none
PointsTo($\mathcal{L}$(pd2))	$\varnothing$	{$\mathcal{L}$(d)}	{$\mathcal{L}$(c), $\mathcal{L}$(d)}

a work queue for newly updated locations and takes one root location from the queue to update all other leaf locations until the queue is empty. Since each root takes $O(N)$ iterations to update others, and one location can be updated and queued at most a constant number of times (the height of the lattice is constant because all properties are either boolean or integer but limited to the scope depth), this algorithm has $O(N^2)$ time complexity.

4.2 Completeness Analysis

Explicit deallocation requires a complete points-to set to ensure that every possibly-pointed-to object is safe to free. That is, the lifetime of an object cannot exceed the pointer pointing to it. If the analysis does not have a complete points-to set for a certain pointer, then freeing this pointer could end up freeing an unexpected object, which may require a longer lifetime, leading to unsoundness.

Go's escape analysis produces a complete points-to set for some locations, but it does not guarantee a complete points-to set for every location due to its simplification on indirect stores. For example, in fig. 1, c's address is held by pd2 due to the indirect store on line 24. c is safely heap-allocated according to the conservative $\langle \mathcal{L}(\texttt{pc}), \texttt{heapLoc} \rangle$ edge, but this does not tell pd2 anything about c. *PointsTo*($\mathcal{L}$(pd2)) lacks c, which means it is incomplete.

Table 3 compares *PointsTo*($\mathcal{L}$(pd2)) in different kinds of escape analysis. The faster the algorithm is, the more data flow information it omits. Only the connection graph always provides complete points-to sets. Whenever encountering dereferencing or field accessing, Fast Escape Analysis gives incomplete points-to sets. Whenever encountering an indirect store, the Go escape graph gives incomplete points-to sets.

The part of the escape graph not affected by indirect stores can contain complete, precise points-to sets. For example, in fig. 1, *PointsTo*($\mathcal{L}$(s)), *PointsTo*($\mathcal{L}$(bigObj)), and *PointsTo*($\mathcal{L}$(pc)) are complete and precise.

To identify which locations are unaffected by indirect stores and thus have complete points-to sets in Go's escape graph, we introduce two new properties for locations and their constraints.

Definition 4.11 (Exposes)**.** *Exposes*(l) is true if untracked modifications to locations in *PointsTo*(l) may have been made by storing indirectly into l. *Exposes*(l) is true if

- l = heapLoc, or
- l = return, or
- $l = \mathcal{L}(n)$ and n is the destination of an indirect store (`*n = ...`), or
- $\exists m \,.\, l \in \mathit{Holds}(m) \wedge \mathit{MinDerefs}(l, m) \leq 0 \wedge \mathit{Exposes}(m)$.

Exposes(l) does not mean l has an incomplete points-to set. It means that l exposes the addresses of locations in *PointsTo*(l) to an under-tracked data flow, so indirect stores might change their values elsewhere and are thus incomplete. For example, in fig. 1, *Exposes*($\mathcal{L}$(pc)) is true because it exposes c's address, but $\mathcal{L}$(pc) itself is complete as all changes to its value are tracked, and it only fetches c's address.

Definition 4.12 (Incomplete)**.** *Incomplete*(l) is true if l can point to locations not in *PointsTo*(l) at run time. *Incomplete*(l) is true if

- l is a formal parameter, or
- $\exists m \,.\, l \in \mathit{PointsTo}(m) \wedge \mathit{Exposes}(m)$, or
- $\exists m \,.\, m \in \mathit{Holds}(l) \wedge \mathit{Incomplete}(m)$.

Exposes() and *Incomplete*() need not be computed for data types not containing pointers, such as scalars, scalar arrays, and objects containing only these types.

GoFree enhances the property propagation algorithm shown in fig. 5 to support newly introduced constraints. The last constraint in definition 4.12 propagates properties from leaf locations to root locations, which is in the reversed direction of previous constraints. Lines 10–13 in fig. 5 support back-propagation. It is still an $O(N^2)$ algorithm. The base Go algorithm updates and re-queues only leaf locations on each walk from the root. GoFree also allows the updating and re-queuing of the root itself. The root still has only a constant number of properties to be updated, so it can only be re-queued at most constant times. So, the $O(N^2)$ complexity is not increased.

As described so far, GoFree can compute a location's points-to set and decide its completeness in $O(N^2)$ time.

4.3 Lifetime Analysis

GoFree models lifetime at a scope accuracy — GoFree determines the smallest (innermost, deepest) scope within a function that contains the definition and all uses of an object. If no such scope exists, the object escapes from the function.

Go's stack allocation optimization determines whether or not an object lives beyond its declaration scope (the innermost brace pair in which it is allocated). Stack allocation has one opportunity to free an object, which is at the end of its scope. By contrast, GoFree's lifetime analysis can find out how many layers of scopes the object has escaped from and determine the exact scope where it is safe to free the object. GoFree can free objects that escape from several scopes

and even the function (with inter-procedural analysis, section 4.4).

Definition 4.13 (DeclDepth). *DeclDepth*(l) $\in \mathbb{Z}$ records the scope depth at the declaration of variable `n`, where $l = \mathcal{L}(\mathtt{n})$. For dummy locations, *DeclDepth*(`heapLoc`) $= -1$, and *DeclDepth*(`return`) $= -1$.

Definition 4.14 (OutermostRef). *OutermostRef*(l) $\in \mathbb{Z}$ is the depth of the smallest (innermost) scope in a function that covers l's lifetime. It takes the greatest value satisfying the following two constraints:

- *OutermostRef*(l) $\leq$ *DeclDepth*(l),
- $\forall m \,.\, l \in \mathit{PointsTo}(m) \Rightarrow \mathit{OutermostRef}(l) \leq \mathit{DeclDepth}(m)$.

OutermostRef's value is always taken from some location's *DeclDepth* and does not further propagate. As a result, despite being an integer, it is capped at the maximum *DeclDepth*, so it will not add complexity to the $O(N^2)$ propagation algorithm.

If a pointer's lifetime ends before the lifetime of the object it points to, then we say the pointer has been "outlived" and is not safe to free.

Definition 4.15 (Outlived). *Outlived*(m) is true if a pointer m has a shorter lifetime than any object that it points to.

- $\mathit{Outlived}(m) = \exists l \,.\, (l \in \mathit{PointsTo}(m) \wedge \mathit{OutermostRef}(l) < \mathit{DeclDepth}(m))$

The definitions of OutermostRef and Outlived are syntactically similar. *OutermostRef*(l) is a property of an object l. It is l answering the question: what are the pointers that point to me, who has the longest lifetime among them, and what is the scope of that longest lifetime? *Outlived*(m) is a property of a pointer m answering the question: what are the objects that I point to, what other pointers also point to them, and is any of these pointers referencing them from a bigger scope than I do? Since the pointer m can also be seen as an object from the viewpoint of a higher-order pointer, pointers have both *OutermostRef* and *Outlived* values. On the contrary, for non-pointers, *Outlived* is not defined.

As explicit deallocation applies only to heap objects, there is no need to call `tcfree` on a pointer that only points to stack objects. (Such a pointer will be stack-allocated.) However, such a call is safe because `tcfree` ignores stack objects.

Definition 4.16 (PointsToHeap). *PointsToHeap*(m) is true if m might point to a heap object.

- $\mathit{PointsToHeap}(m) = \exists l \,.\, l \in \mathit{PointsTo}(m) \wedge \mathit{HeapAlloc}(l)$.

The final step is to determine which locations to free using `tcfree`. This location must be both safe and worthy of free.

Definition 4.17 (ToFree). *ToFree*(m) is true if m is qualified to be deallocated by `tcfree`.

$\mathit{ToFree}(m) = \neg\mathit{Incomplete}(m) \wedge \neg\mathit{Outlived}(m) \wedge \mathit{PointsToHeap}(m)$

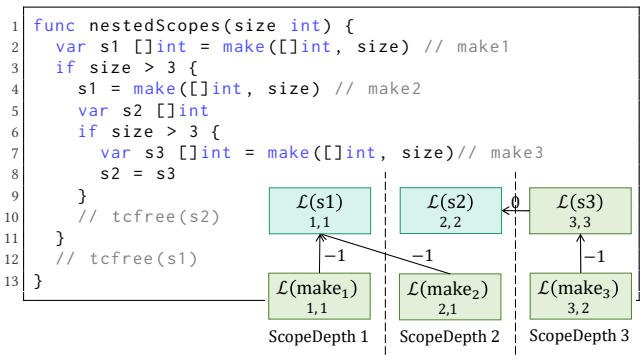

Figure 6: Nested scopes. Numbers under each node are their *DeclDepth*() and *OutermostRef*() values.

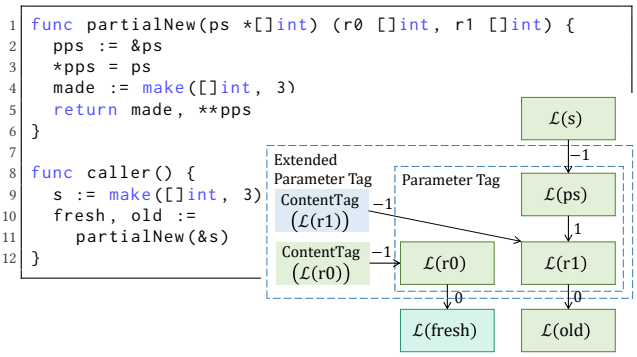

Figure 7: Inter-procedural analysis. Content tags summarize the escape properties of what a return value points to.

The nested scopes in fig. 6 illustrate lifetime analysis. All three slices in fig. 6 are heap-allocated because they have non-constant size. The lifetime analysis identifies that both `s1` and `s2` have complete points-to sets and are not outlived by the objects they point to. Therefore, two `tcfrees` are inserted to free them. However, `s3` has passed the address of its underlying object to an outer scope, making it outlived, so it is not safe to free it within its scope.

4.4 Inter-procedural Analysis

Go's inter-procedural escape analysis computes function summaries called *parameter tags*. Each parameter tag is a compressed escape graph of the function: a bipartite graph with edges from parameters to returns. A parameter tag's locations include only the function's parameter and return values. The detailed data flow within the function is compressed into edges directly from parameters to return values. The edges' *Derefs*() are taken from *MinDerefs*() on the full escape graph.

At a call site, a copy of the callee's parameter tag is embedded into the caller as a subgraph. Go orders the intra-procedural analysis of functions inner-to-outer so that more call sites will find known parameter tags. If a callee's parameter tag is unknown (possibly due to recursion or closures), Go uses a default parameter tag where all parameters flow to the heap and all return values come from the heap.

Parameter tagging does not support explicit deallocation because it does not include information about what new allocations that the return values bring out. As shown in fig. 7, `fresh` does not see the inner `make` from $PointsTo(\mathcal{L}(\texttt{r0}))$, so Go misses the opportunity to deallocate it.

GoFree uses a new approach. *Content tagging* summarizes return values' points-to sets so that newly allocated objects in the callee can be explicitly deallocated in the caller. For each return value location l, GoFree adds a dummy *content tag* location $m = ContentTag(l)$ to summarize its points-to set, and an edge $e = \langle m, l \rangle$ with $Derefs(e) = -1$. After intra-procedural analysis, GoFree adjusts a few of l's and m's properties before adding them both to the extended parameter tag:

- $LoopDepth(l) = +\infty$
  $DeclDepth(l) = +\infty$
  $LoopDepth(m) = +\infty$
  $DeclDepth(m) = +\infty$
- $HeapAlloc(m) = PointsToHeap(l)$
- $Incomplete(l) = Incomplete(m)$

The first rule sets the depths to a large enough value so that they do not appear in the caller as if used by an outer scope. The second rule passes into $HeapAlloc(m)$ whether any location in $PointsTo(l)$ is heap allocated. The third rule tells l to use the incompleteness property of m rather than its own. This is because $Incomplete(l)$ may be propagated from a formal parameter of the callee. As we have conservatively set $Incomplete(param) = \texttt{true}$ in case we have no information about the caller, this $Incomplete(param) = \texttt{true}$ modeling may now cause a false positive now that we have information from the caller in inter-procedural analysis. In contrast, $Incomplete(m)$ could only come from indirect stores within the callee, which must be recorded for safety.

As shown in fig. 7, $ContentTag(\mathcal{L}(\texttt{r0}))$ propagates a deallocation opportunity to `fresh`, which improves analysis accuracy. $ContentTag(\mathcal{L}(\texttt{r1}))$ propagates its incompleteness to $\mathcal{L}(\texttt{old})$, so it will not be freed because the existence of an indirect store inside the callee can be obtained from the caller via the tag.

4.5 `tcfree` Instrumentation

GoFree's runtime includes APIs shown in table 4 that deallocate objects of different types and sizes. Figure 8 shows their calling relationships.

`TcfreeSlice` and `TcfreeMap` are specialized variants of `tcfree` that deallocate built-in data structures of Go (section 4.6).

`Tcfree` receives an address, either from a raw pointer or an unwrapped slice or map, and forwards the address to `TcfreeSmall` or `TcfreeLarge` according to its size.

`TcfreeSmall` and `TcfreeLarge` adopt different strategies to deallocate objects of different sizes efficiently. For their implementation, see section 5.

For each location whose *ToFree*() is `true`, the GoFree compiler inserts a corresponding variant of `tcfree` as the last statement (except for `return` and `goto`, so the `tcfree` is live) of its declaration scope. If its lifetime scope is smaller than its declaration scope, GoFree's analysis is not accurate enough to detect that. In most cases, the program executes to the last line of a scope, so `tcfree` will be reached. In cases where the function returns from the middle, `tcfree` is not reached, but it is still safe to leave the deallocation to GC.

4.6 Support for Go Language Features

4.6.1 Slice. Slice is a built-in list implementation in some modern languages, such as Go, Python, and Rust. A slice is typically a fat pointer composed of the address of the underlying array and its length and capacity.

A slice is challenging for escape analysis because its memory is runtime-managed and does not behave as an ordinary object. In an escape graph, slices are equivalent to pointers to their underlying arrays, but these arrays are not always explicitly allocated. When appending to a slice that does not already have an underlying array, one is implicitly allocated. When appending to a full slice, the runtime reallocates a larger space to extend it. If these implicit allocations and data flows are not reflected in the escape graph, they can cause missing optimizations or even safety problems.

GoFree supports slices by adding a dummy content location m and setting $HeapAlloc(m) = \texttt{true}$ upon each slice appending, modeling a possible implicit allocation. GoFree connects it to the slice location l with edge $e = m \xrightarrow{-1} l$. Slice appending in a loop usually causes heap allocation, which introduces explicit deallocation opportunities.

Slices also provide great opportunities for GoFree optimization. They are usually large and do not have a fixed size, making them hard to stack-allocate. Slices are used a lot as temporary buffers with relatively simple dataflow, usually local to a scope. GoFree has a variant of `tcfree` for slices called `TcfreeSlice`. When a slice is passed to it, this runtime API unwraps the address of its underlying array and forwards it to deallocation implementations.

4.6.2 Map. Map is Go's built-in implementation of the hash table. The Go runtime manages its memory. Every map access is compiled into a runtime call, which may cause allocation. The biggest part of a map is its *buckets*, a continuous array. When the map's load factor reaches a certain constant,

Table 4: The `tcfree` family. Parameters are always addresses.

Runtime API	Parameter	Functionality
`TcfreeSlice`	slice	unwraps the address of underlying array, calls `Tcfree`
`TcfreeMap`	map	unwraps the address of underlying buckets, calls `Tcfree`
`Tcfree`	object	calls `TcfreeSmall` or `TcfreeLarge`
`TcfreeSmall`	small obj.	deallocates a small object from mcache
`TcfreeLarge`	large obj.	deallocates a large object from mcentral

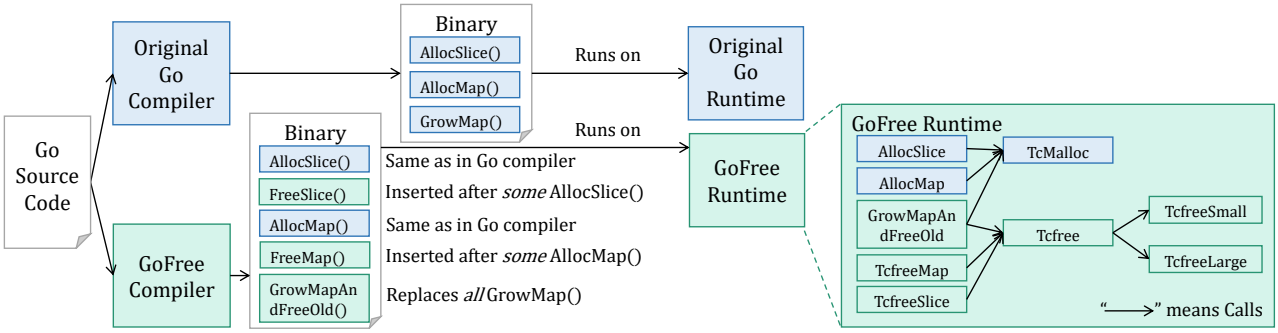

Figure 8: `tcfree` components. The binary generated by the compiler contains calls to runtime routines such as AllocSlice. Blue boxes are in the original Go system. Teal boxes are new to GoFree.

a bigger new bucket array is allocated in the heap to replace the old one.

Go maps provide GoFree two optimization opportunities: when they grow and when ends its lifetime.

First, when a map grows, its old bucket array is evacuated and abandoned. Since different maps do not share the same bucket array, this abandoned array is in the growing map's exclusive ownership. GoFree deallocates it. This is a runtime optimization that needs no static analysis but does need `tcfree`. We observe that, even in a managed language, an explicit free primitive can improve memory efficiency.

Second, a map holds an underlying bucket array when its lifetime ends. GoFree detects this case similarly to slices. It uses a specialized `TcfreeMap` primitive to unwrap the address of a map's underlying bucket array and forward it to deallocation implementations.

4.6.3 Multiple Return Values.

Go provides native language support for multiple return values. Unlike Python or Modern C++, who wrap return values into a tuple and provide a syntactic sugar to simulate multiple return values, Go regards each return value as an independent object.

Previous allocation analysis for single-return-value functions typically identifies factory methods [19] where the object returned is newly allocated by the callee. The analysis treats a factory method as a `new` expression in the caller.

Unfortunately, this technique does not work for Go. Each Go function may return many objects (fig. 7), both newly allocated ones $\mathcal{L}(\texttt{r0})$ and received ones $\mathcal{L}(\texttt{r1})$. A Go function may be a factory concerning some return values but not concerning others.

GoFree's inter-procedural analysis generalizes factory method identification with extended parameter tagging and provides enough information to handle multiple return values precisely. Essentially, a factory method is a function that returns something worthy of freeing. GoFree records the *PointsToHeap* property of each return value into the *HeapAlloc* property of its content tag. GoFree also records the completeness property for safety.

4.6.4 Function Inlining.

Function inlining is a common compiler optimization to reduce the cost of calling short functions. Inlined functions are embedded into the caller as part of its code, reducing expensive frame maintenance operations.

Go's escape analysis benefits from inlining. Go stack allocation does not track an object once it escapes from the function and heap allocates it. However, if the callee is inlined, more scopes in the caller are visible to the callee object. If an object escapes from the inlined callee but not from the caller, Go can still allocate it on the stack.

GoFree does not depend on inlining because its inter-procedural analysis provides enough information to analyze the caller function as precisely as the intra-procedural analysis does. Our extended parameter tags model dataflow from parameters to return values, and provide both *HeapAlloc*() and *Incomplete*() properties for return values to analyze them safely and precisely.

5 Runtime Support for `tcfree`

Figure 8 illustrates the components of the original Go compiler and our GoFree compiler. Go programs that use built-in data structures like slices and maps are compiled into executable binaries containing corresponding runtime calls for object allocation (and explicit deallocation, for GoFree). Type-specific primitives are inserted into user programs to support different built-in data structures. Calls to them are classified by `tcfree` and routed to `TcfreeSmall` or `TcfreeLarge` depending on the size of the deallocated object.

`tcfree` includes a family of explicit free primitives targeting objects of various types and sizes. It aims to enable safe and low-cost explicit deallocation of heap objects. The `tcfree` primitive prioritizes safety and efficiency; it does not guarantee successful explicit deallocation.

The deallocation strategy is tightly associated with Go's allocation strategy as implemented by TCMalloc, which is dependent on the size of objects.

For all implementations in the `tcfree` family, `tcfree` gives up deallocating and returns without making any changes in the following two cases. When GC has been triggered and is already running concurrently with the user program, `tcfree`

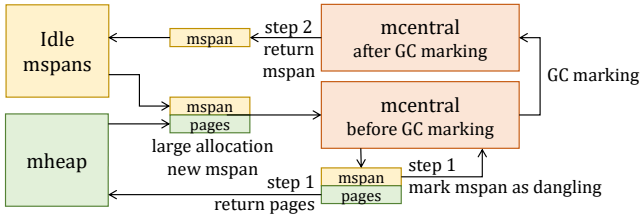

Figure 9: `tcfree` for large objects is implemented with a 2-step strategy. Yellow blocks are mspan control blocks. Green blocks are memory pages where large objects reside.

does not race GC to deallocate any object. If an mspan's ownership has changed due to Go's runtime thread scheduling between allocation and `tcfree`, `tcfree` gives up because it is unsafe to operate on another thread's mspan. There are also other implementation-specific situations where `tcfree` gives up, to be introduced in corresponding paragraphs.

Explicit Deallocation of Small Objects. Most heap objects are allocated in thread-local mspans, enabling efficient deallocation without locking. If an object of a given size fits in an mspan, TCMalloc allocates it lock-free. This is accomplished by bumping the mspan's free index and setting an allocation bit. In most cases, `tcfree` is called on a short-lived object, which means that the object is likely still located in the same mspan. If so, deallocation is performed by simply reverting the free index and clearing the allocation bit. If the mspan has been filled and swapped out, `tcfree` does not risk the safety and efficiency of the deallocation process; it just returns without changing anything.

Explicit Deallocation of Large Objects. When objects are too large to fit into a thread-local mspan, TCMalloc creates a new mspan exclusively for the object. This involves taking an mspan struct and associating several memory pages to it (fig. 9). After allocation, the mspan is pushed into the mcentral, the central cache for mspans not owned by any single thread. `tcfree` works in 2 steps to minimize its effect on GC (fig. 9). In the first step, `tcfree` locks the mspan and returns the memory pages it owns to mheap, the central cache for unused memory pages. `tcfree` also marks the mspan dangling in the step. When the marking mechanism of GC encounters a dangling mspan, GC skips the mspan instead of marking it and its descendants. In the second step, the mspan is returned to a global list of idling mspans just like any other mspans that are not marked after a GC marking phase.

Double-free Handing. Under the current GoFree implementation, a double-free will occur if an object is pointed to by more than one pointer, all of which are in the same scope and eligible for `tcfree`. However, this will not cause trouble because (1) `tcfree` will ignore any already-freed memory, (2) these `tcfree`s are inserted adjacently and no preemption or new allocation can happen in between, and (3) in TCMalloc's policy, no other thread can access the part of memory cached by the current thread.

Possibility of Batching. Batching `tcfree`s of different objects in the same scope is possible. It can reduce the overhead of `tcfree` calls, but typically offers limited performance gains since few objects are freed in a single scope.

Batching frees at run time is also possible but offers little gain. Most of the cost inside `tcfree` comes from checking the object's and its mspan's status to determine whether it is safe to free, which needs to be done right before freeing regardless of batching. Also, delaying `tcfree`s can increase the likelihood of `tcfree` returning doing nothing due to objects having been swapped out of the current thread's cache.

Safety upon Defer() and Panic(). `tcfree` handles `Defer()` and `Panic()` calls conservatively. Any object passed to a `Defer()` or `Panic()` call is banned from freeing.

No panic can be put between `tcfree`s due to the way they are inserted by GoFree. If a panic is called before several `tcfree`s that are in a row, deallocation does not happen and is left to GC. Signals and interrupts are OS-level control flow changes, not Go runtime control flow changes. They might execute some OS code between two `tcfree`s (or even inside a `tcfree`), but they do not lead to Go runtime calls, so it's safe.

6 Evaluation

This section evaluates six aspects of the proposed approach.

1. *Microbenchmark on effectiveness.* How does the size of explicitly deallocated objects affect run time and heap memory occupation?
2. *Real-world performance improvement.* How well does GoFree perform on real-world applications?
3. *Deallocation target selection.* Why does GoFree choose to free only slices and maps?
4. *Contribution breakdown.* How much does each category of objects contribute to explicitly deallocated space?
5. *Compilation speed.* Does GoFree retain fast compilation?
6. *Robustness.* Does our GoFree implementation corrupt memory?

6.1 Experimental Setup

All experiments were conducted on an Ubuntu server with two Intel(R) Xeon(R) Gold 6248R CPUs (each has 48 cores (96 threads) and runs at 3.00GHz) and 503GB memory. A single Go process is allowed to create up to 32 threads.

We implemented GoFree by modifying the official Go 1.17.7 compiler's source code. We modified fewer than 2000 lines of non-blank non-comment lines of code (not including profiling utilities).

Table 5 describes our metrics. Our profiling tool is implemented by hooking Go's runtime library and collecting information throughout the program's execution. The metrics are collected upon specific runtime calls and events, such

Table 5: Experimental metrics

Metric	Description
time	wall-clock time of one program execution
GC Time	time minus time of a run of the same program with GC turned off
GCs	count of GC cycles triggered in one execution
alloced	allocation size: total amount of heap allocation in one execution
freed	free size: total heap memory freed by `tcfree` in one execution
free ratio	free ratio = freed / alloced
maxheap	maximum heap size during one execution

Table 6: Descriptions of subject open-source Go projects

Project	Description
Go	The Go compiler
hugo	A webpage generator that converts markdown into HTML
badger	A key-value database
Go/json	The standard Go library for JSON parsing and manipulation
scheck	A static checking tool for Go from `dominikh/go-tools`
slayout	Struct layout analysis tool from `dominikh/go-tools`

```
1 func mapPopulate() {
2   for n := 0; n < 10000000/C; n++ {
3     m := make(map[int]int)
4     for i := 0; i < C; i++ {
5       m[i] = i
6     }
7   }
8 }
```

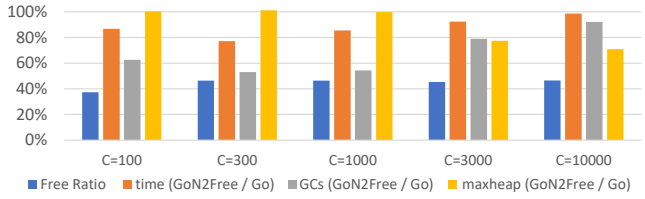

Figure 10: Microbenchmark map experiment. A bigger `C` value (used in the code above) means that the average size of the deallocated objects is bigger.

as heap allocation, `tcfree`, and triggering of a new GC cycle. Our profiling tool has no observable effect on the program's performance.

6.2 Subject Programs

The Go community lacks a widely accepted and consistently used benchmark suite. We chose 6 open-source programs (table 6) from GitHub. Each one had more than 5k stars and was a program rather than a service. (It is harder to fairly measure the run time and memory consumption of a service.)

6.3 Microbenchmark on Effectiveness

Go's TCMalloc is a complex memory allocator with multiple levels of caches. The effect of deallocation goes through these cache layers and the Go GC mechanism before affecting overall performance. Figure 10 shows the results of a microbenchmark experiment. Explicit deallocation may reduce time (orange bar), memory (yellow bar), or both. Which of the two benefits more depends on the average size of deallocated objects.

Table 7: Effect of GoFree's optimizations. Data in grey are not significant at $p = 0.01$, i.e., neither GoFree nor Go is better. Ratios (%) are GoFree / Go. Values less than 100% indicate that GoFree is better.

Project	time			GC time	GCs			free	maxheap		
	ratio	stdev	p-value	ratio	ratio	stdev	p-value	ratio	ratio	stdev	p-value
Go	99%	2%	< 0.001	92%	95%	3%	< 0.001	12%	99%	3%	< 0.001
hugo	100%	11%	0.460	100%	98%	2%	< 0.001	6%	99%	7%	0.288
badger	100%	3%	0.253	98%	95%	9%	< 0.001	4%	100%	9%	0.420
json	94%	1%	< 0.001	55%	77%	0%	< 0.001	23%	96%	0%	< 0.001
scheck	98%	2%	< 0.001	83%	94%	4%	< 0.001	15%	96%	5%	< 0.001
slayout	99%	6%	0.098	5%	98%	3%	< 0.001	25%	89%	2%	< 0.001
average	98%	N/A	N/A	87%	93%	N/A	N/A	14%	96%	N/A	N/A

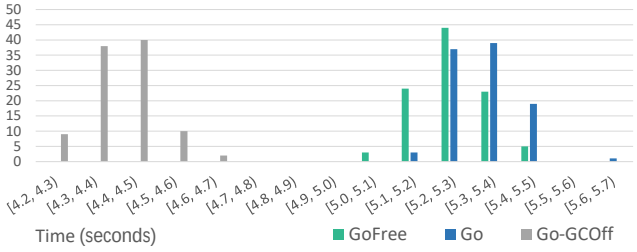

Figure 11: Run time distribution across 99 runs under each of the 3 settings respectively: original GoFree, original Go, and original Go with GC turned off.

As shown in fig. 10, each value of `C` results in a similar free ratio (blue bar), indicating comparable amounts of deallocation. However, as `C` grows larger, the mean size of deallocated objects increases, resulting in greater reductions in heap size and less significant reductions in GC frequency (grey bar) and run time.

6.4 Real-world Performance Improvement

Results are in table 7. Our metrics behave as a random distribution across different runs (fig. 11 is typical), so we ran each program under 3 settings, 99 times for each setting, and take the average. The settings are: (1) compile with Go, (2) compile with GoFree, and (3) compile with Go but turn off GC at run time (Go-GCOff). We compare time, GC and maxheap between settings (1) and (2) to get the "ratio" columns in table 7. GC time ratio is $(time_{GoFree} - time_{GoGCOff})/(time_{Go} - time_{GoGCOff})$.

Go provides no way to set a strict heap size limit. The "`GOGC`" environment variable sets a soft goal for the Go GC pacing mechanism, indicating the desired heap growth percentage relative to the heap size at the end of the last GC cycle before triggering the next GC cycle. Go will never trade a stop-the-world GC for enforcing this soft goal. More recent versions of Go than 1.17.7 (which we used) do provide a way to set a memory limit, but that is also a similar soft goal mechanism.

On average of the 6 programs, GoFree deallocated 14% of allocated heap memory, reduced GC frequency by 7%, GC time overhead by 13%, wall-clock time by 2%, and heap size by 4%.

Table 8: Stack/heap allocation decisions of slices, maps, and other data structures.

Project	Stack others	Heap GC others	Stack slices	Heap tcfree slices	Heap GC slices	tcfree/ (tcfree + GC)	Stack maps	Heap tcfree maps	Heap GC maps	tcfree/ (tcfree + GC)
Go	461583	1021	2288	660	4692	12%	142	304	431	41%
hugo	476897	3698	2521	728	7314	9%	373	159	1271	11%
badger	145569	446	1323	156	2234	7%	89	24	100	19%
Go/json	45244	45	241	34	280	11%	4	3	1	75%
scheck	166269	455	1123	263	2417	10%	77	84	201	30%
slayout	78249	167	364	94	822	10%	22	32	84	28%
average	N/A	N/A	N/A	N/A	N/A	10%	N/A	N/A	N/A	34%

A wall-clock speedup of 2% is not dramatic. However, at scale (such as in a data center) it is important; it is in line with compiler optimizations; and it comes with no drawbacks. A referee remarked that our experiments likely underestimated the benefits of the technique, because we did not evaluate on Go services with high queries per second that create a lot of objects per query. That is common in practice (and GC can affect tail latency, which is more important than overall execution time), but we have no access to proprietary code with those characteristics.

In statistically significant items (a metric for a project), GoFree always reduces GC frequency, sometimes reduces time and maxheap, and never increases time or maxheap. GoFree reduces most time from `json`, by 6% (ratio = 94%), and most maxheap from `structlayout`, by 101% (ratio = 89%). Typically, programs with higher free ratio benefit more from GoFree. Apart from the programs listed, we also briefly tested on `protobuf-go`, `fastjson`, `fzf`, `gods`, and benchmarks in Sweet (version Jul.5, 2024) [20] (except for the Go compiler which is also included in Sweet). They have too low free ratio (< 5%) so we assume GoFree will not have a significant effect.

The Go compiler is written in Go, so it can benefit from GoFree. When GoFree is compiled by GoFree, it compiles other Go programs faster by 1%.

6.5 Deallocation Target Selection

Table 8 reports how often Go stack allocates slices, maps, and all other data structures including user-defined ones. Columns 2 and 3 show that Go's stack allocation is very effective for the "other" category. Based on this, GoFree operates only on stacks and maps, to avoid adding overhead but little benefit for other types. Other columns of table 8 show that GoFree achieves significant savings for slices and maps.

6.6 Contribution Breakdown

`tcfree` can reclaim memory from three sources: when a slice's lifetime ends, when a map's lifetime ends, and the old bucket when a map grows. Their contribution depends on the application (table 9). For example, the Go compiler uses a lot of slices to hold basic blocks temporarily during compilation, so it benefits a lot from `FreeSlice()`. Applications that do not use a certain data structure cannot benefit from it.

Table 9: Contribution breakdown of total space reclaimed by the three deallocation categories. Each row sums to 100%.

Project	`FreeSlice()`	`FreeMap()`	`GrowMapAndFreeOld()`
Go	56%	14%	30%
hugo	56%	14%	30%
badger	0%	0%	100%
json	0%	0%	100%
scheck	2%	50%	48%
slayout	1%	0%	99%

6.7 Compilation Speed

To show that GoFree meets its design goal of maintaining Go's compilation speed, we compiled the `ssa` package, a relatively large package of the Go compiler, 99 times with the original Go compiler and GoFree. The values are similar and the p-value is 0.496, meaning the difference is insignificant.

6.8 Robustness

To test whether the GoFree implementation corrupts the memory i.e., deallocates any living object, we ran Go's official package tests using a mock implementation of `tcfree`. Instead of deallocating the memory according to the strategies described in section 5, this mock implementation sets the memory to zero, or flips all the bits. This reveals wrong deallocation faster because it causes any later usage of the freed space to get an incorrect value instead of possibly getting the correct value if the slot is not immediately re-allocated. In multiple runs, GoFree with the mock `tcfree` implementation successfully passed all the tests. This suggests that the GoFree algorithm is safe.

7 Conclusion

GoFree is a new $O(N^2)$ static analysis that identifies short-lived heap objects that are capable of bypassing GC via explicit freeing, without increasing the time complexity of Go's escape analysis. The approach identifies complete points-to sets and performs lifetime analysis upon them. Key ideas include offloading complexity from static analysis to the runtime system, allowing tentative free primitives for optimization and concurrency, and identifying when a conservative analysis performs precisely. These ideas are applicable to any language that supports garbage collection and either region allocation or explicit deallocation.

Acknowledgments

This research was partly funded by the National Natural Science Foundation of China (Grant No. 62272434). We thank Yigong Hu for sharing insights on experiment design. We also appreciate the constructive feedback from reviewers at CGO and other conferences we have submitted to, which greatly contributed to its development.

References

[1] Alex Aiken, Manuel Fähndrich, and Raph Levien. 1995. Better Static Memory Management: Improving Region-Based Analysis of Higher-Order Languages. In *PLDI '95: Proceedings of the SIGPLAN '95 Conference on Programming Language Design and Implementation*. La Jolla, CA, USA, 174–185.

[2] Lars Ole Andersen. 1994. *Program analysis and specialization for the C programming language*. Ph. D. Dissertation. Citeseer.

[3] Richard Bellman. 1958. On a routing problem. *Quarterly of applied mathematics* 16, 1 (1958), 87–90.

[4] Daniil Berezun and Dmitri Boulytchev. 2014. Precise garbage collection for C++ with a non-cooperative compiler. In *CEE-SECR 2014: Proceedings of the 10th Central and Eastern European Software Engineering Conference in Russia*. Moscow, Russia, Article 15.

[5] Bruno Blanchet. 1999. Escape analysis for object-oriented languages: application to Java. In *Proceedings of the 14th ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications* (Denver, Colorado, USA). 20–34.

[6] Chandrasekhar Boyapati, Alexandru Salcianu, William Beebee, Jr., and Martin Rinard. 2003. Ownership types for safe region-based memory management in real-time Java. In *PLDI 2003: Proceedings of the ACM SIGPLAN 2003 Conference on Programming Language Design and Implementation*. San Diego, CA, USA, 324–337.

[7] Steve Carr and Ken Kennedy. 1994. Scalar replacement in the presence of conditional control flow. *Software: Practice and Experience* 24, 1 (1994), 51–77.

[8] Hao Chen and Hee-Jong Suh. 2012. An improved Bellman-Ford algorithm based on SPFA. *The Journal of the Korea institute of electronic communication sciences* 7, 4 (2012), 721–726.

[9] Sigmund Cherem and Radu Rugina. 2007. Uniqueness inference for compile-time object deallocation. In *Proceedings of the 6th International Symposium on Memory Management*. 117–128.

[10] Jong-Deok Choi, Manish Gupta, Mauricio Serrano, Vugranam C Sreedhar, and Sam Midkiff. 1999. Escape analysis for Java. In *Proceedings of the 14th ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications* (Denver, Colorado, USA). 1–19.

[11] Jong-Deok Choi, Manish Gupta, Mauricio J Serrano, Vugranam C Sreedhar, and Samuel P Midkiff. 2003. Stack allocation and synchronization optimizations for Java using escape analysis. *ACM Transactions on Programming Languages and Systems* 25, 6 (2003), 876–910.

[12] Matthew Davis, Peter Schachte, Zoltan Somogyi, and Harald Søndergaard. 2013. A low overhead method for recovering unused memory inside regions. In *MSPC 2013: Proceedings of the ACM SIGPLAN Workshop on Memory Systems Performance and Correctness*. Seattle, WA, USA, Article 4.

[13] Ulan Degenbaev, Jochen Eisinger, Kentaro Hara, Marcel Hlopko, Michael Lippautz, and Hannes Payer. 2018. Cross-component garbage collection. In *OOPSLA 2018, Object-Oriented Programming Systems, Languages, and Applications*. Boston, MA, USA, Article 151.

[14] Ulan Degenbaev, Michael Lippautz, and Hannes Payer. 2019. Garbage collection as a joint venture. *CACM* 62, 6 (June 2019), 36–41.

[15] Alain Deutsch. 1997. On the complexity of escape analysis. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. 358–371.

[16] L Peter Deutsch and Daniel G Bobrow. 1976. An efficient, incremental, automatic garbage collector. *Commun. ACM* 19, 9 (1976), 522–526.

[17] Fanding Duan. 1994. A faster algorithm for shortest-path—SPFA. *Journal of Southwest Jiaotong University* 29, 2 (1994), 207–212.

[18] Martin Elsman. 2003. Garbage collection safety for region-based memory management. In *TLDI 2003: The ACM SIGPLAN Workshop on Types in Language Design and Implementation*. New Orleans, LA, USA, 123–134.

[19] David Gay and Bjarne Steensgaard. 2000. Fast escape analysis and stack allocation for object-based programs. In *Proceedings of the International Conference on Compiler Construction*. Springer, 82–93.

[20] Go. 2024. Sweet: Benchmarking suite for Go implementations. https://pkg.go.dev/golang.org/x/benchmarks/sweet.

[21] Google. 2022. Go Escape Analysis Invariants. https://github.com/golang/go/blob/c5adb8216968be46bd11f7b7360a7c8bde1258d9/src/cmd/compile/internal/escape/escape.go#L23.

[22] Google. 2022. Go Official Frequently Asked Questions. https://go.dev/doc/faq#stack_or_heap.

[23] Google. 2023. The Go Programming Language Specification. https://go.dev/ref/spec.

[24] Google. 2023. A Guide to the Go Garbage Collector. https://go.dev/doc/gc-guide.

[25] Dan Grossman, Greg Morrisett, Trevor Jim, Michael Hicks, Yanling Wang, and James Cheney. 2002. Region-Based Memory Management in Cyclone. In *PLDI 2002: Proceedings of the ACM SIGPLAN 2002 Conference on Programming Language Design and Implementation*. Berlin, Germany, 282–293.

[26] Samuel Z. Guyer, Kathryn S. McKinley, and Daniel Frampton. 2006. Free-Me: A Static Analysis for Automatic Individual Object Reclamation. In *Proceedings of the 27th ACM SIGPLAN Conference on Programming Language Design and Implementation* (Ottawa, Ontario, Canada). ACM, New York, NY, USA, 364–375. https://doi.org/10.1145/1133981.1134024

[27] Niels Hallenberg, Martin Elsman, and Mads Tofte. 2002. Combining region inference and garbage collection. In *PLDI 2002: Proceedings of the ACM SIGPLAN 2002 Conference on Programming Language Design and Implementation*. Berlin, Germany, 141–152.

[28] Matthew Hertz and Emery D. Berger. 2005. Quantifying the performance of garbage collection vs. explicit memory management. In *OOPSLA 2005, Object-Oriented Programming Systems, Languages, and Applications*. San Diego, CA, USA, 313–326.

[29] Jaehwang Jung, Janggun Lee, Jaemin Choi, Jaewoo Kim, Sunho Park, and Jeehoon Kang. 2023. Modular verification of safe memory reclamation in concurrent separation logic. In *OOPSLA 2023, Object-Oriented Programming Systems, Languages, and Applications*. Cascais, Portugal, Article 251.

[30] William Kennedy. 2018. Escape-Analysis Flaws. https://www.ardanlabs.com/blog/2018/01/escape-analysis-flaws.html.

[31] Kyungwoo Lee and Samuel P Midkiff. 2006. A two-phase escape analysis for parallel Java programs. In *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*. IEEE, 53–62.

[32] Maged M Michael. 2004. Hazard Pointers: Safe Memory Reclamation for Lock-Free Objects. *TPDS* 15, 6 (June 2004), 491–504.

[33] Jon Rafkind, Adam Wick, John Regehr, and Matthew Flatt. 2009. Precise garbage collection for C. In *ISMM 2009: International Symposium on Memory Management*. Dublin, Ireland, 39–48.

[34] Paul Menage Sanjay Ghemawat. 2007. TCMalloc: Thread-Caching Malloc. http://goog-perftools.sourceforge.net/doc/tcmalloc.html.

[35] Kunal Sareen and Stephen Michael Blackburn. 2022. Better understanding the costs and benefits of automatic memory management. In *MPLR 2022: 19th International Conference on Managed Programming Languages & Runtimes*. Brussels, Belgium, 29–44.

[36] Ran Shaham, Eran Yahav, Elliot K Kolodner, and Mooly Sagiv. 2003. Establishing local temporal heap safety properties with applications to compile-time memory management. In *Proceedings of the International Static Analysis Symposium*. Springer, 483–503.

[37] Lukas Stadler, Thomas Würthinger, and Hanspeter Mössenböck. 2014. Partial escape analysis and scalar replacement for Java. In *Proceedings of the IEEE/ACM International Symposium on Code Generation and Optimization*. 165–174.

[38] Mads Tofte and Jean-Pierre Talpin. 1994. Implementation of the typed call-by-value λ-calculus using a stack of regions. In *POPL '94: Proceedings of the 21st Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. Portland, OR, 188–201.

[39] Frédéric Vivien and Martin Rinard. 2001. Incrementalized pointer and escape analysis. In *Proceedings of the ACM SIGPLAN 2001 conference on Programming language design and implementation*. 35–46.

[40] Cong Wang, Mingrui Zhang, Yu Jiang, Huafeng Zhang, Zhenchang Xing, and Ming Gu. 2020. Escape from escape analysis of Golang. In *Proceedings of the 42nd Proceedings of the International Conference on Software Engineering: Software Engineering in Practice*. 142–151.

[41] John Whaley and Martin Rinard. 1999. Compositional pointer and escape analysis for Java programs. In *Proceedings of the 14th ACM SIGPLAN conference on Object-oriented Programming, Systems, Languages, and Applications*. 187–206.

[42] Yu Zhang, Lina Yuan, Tingpeng Wu, Wen Peng, and Quanlong Li. 2010. Just-in-Time Compiler Assisted Object Reclamation and Space Reuse. In *Proceedings of the 2010 IFIP International Conference on Network and Parallel Computing*. Springer-Verlag, Berlin, Heidelberg, 18–34.

[43] Benjamin Zorn. 1993. The measured cost of conservative garbage collection. *Software: Practice and Experience* 23, 7 (1993), 733–756.

Received 2024-05-30; accepted 2024-07-22